Does Workplace Competition Increase Labor Supply? Evidence from a Field Experiment

Amalia R. Miller
University of Virginia
IZA and NBER

Ragan Petrie
Texas A&M University
and Melbourne Institute

Carmit Segal
University of Zurich

July 2, 2019

Abstract

This paper develops a novel field experiment to test the implicit prediction of tournament theory that competition increases work time and can therefore contribute to the long work hours required in elite occupations. A majority of workers in the treatment without explicit financial incentives worked past the minimum time, but awarding a tournament prize increased work time and effort by over 80% and lowered costs of effort or output by over a third. Effort was similar with alternative (piece rate, low-prize tournament) bonuses. Men worked longer than women in the high-prize tournament, but for the same duration in other treatments.

JEL Codes: M52, M55, J16, J22, J33, J44, D91

Keywords: tournaments, performance pay, long work hours, elite occupations, gender

We thank seminar and conference participants at Cornell, HECER Helsinki, WZB Berlin, Colorado State University, UCLA, and the American Economic Association Meetings for helpful comments, and Cailin Slattery and Elliott Isaac for outstanding research assistance. We acknowledge financial support from the Bankard Fund for Political Economy at the University of Virginia.

1. Introduction

Devoting long hours to work is costly to individuals, who must forgo more and more leisure time and home production. The costs are especially onerous for workers with time commitments outside of the labor market, such as workers with family caretaking responsibilities.[1] Yet long hours are common across a range of occupations and are typically required of workers in elite professional careers who compete against one another for pay and promotions (Goldin, 2014; Gicheva, 2013; Lazear, 2018). This paper develops and tests the hypothesis that competition contributes to long work hours, drawing on the prediction from tournament theory that rewarding workers based on their relative performance can induce them to supply high effort (Lazear and Rosen, 1981). Because effort can be increased along the intensive margin, by working harder per unit time, or along the extensive margin, by working longer, an implicit feature of the theory is that competition itself can result in long working hours.

This prediction is intuitive, as it corresponds to popular notions of a "rat race" at work, but it has not previously been tested. We do this by measuring the effects of workplace competition on labor supply and firm costs using a field experiment in which workers are assigned to competitive and non-competitive payment schemes. Our research design enables us to rule out variation due to productive technology because workers are all hired to perform the same job under identical conditions. The only difference is compensation, where workers in the competitive scheme compete for a bonus prize. We also shut the channel of worker self-selection into competition by assigning workers to schemes and by not informing them in advance of the possibility of a bonus prize.

In our implementation, workers were all undergraduate students, offered a fixed ($25) payment for an hour-long research assistance (RA) work session in which they and other students tested and benchmarked a tablet-computer program for a professor. After they arrived, they were assigned to gender-balanced rooms of 4 workers and provided a brief training session. In it, they were told they only needed to work for 10 minutes and then complete a survey about the program to be paid the $25 wage. The nature and purpose of the work, including the value to the employer of the work and of additional effort from workers, were also explained to them.

[1] Women still devote significantly more time to childcare than men do, but paternal time has been increasing in recent decades; men increasingly describe parenthood as essential to their identities and report they enjoy spending time with their children (https://www.pewresearch.org/fact-tank/2018/06/13/fathers-day-facts/).

Workers were asked to try as hard as they could and to stay for as long as they could, for a maximum of 40 minutes. In rooms with tournament pay, the bonus scheme was described to workers at the end of training.

Relative to the experimental literature on tournaments, the primary innovation of our design is that we explicitly allow workers to select both the total duration of their job and the level of effort they exert per unit of time. Previous studies have used a "stated effort" framework (Bull et al., 1987) or focused on the effort intensity margin, either measuring total output for a fixed amount of time (Gneezy et al., 2003; Freeman and Gelber, 2010, Dohmen and Falk, 2011) or speed to complete a fixed task, such as a race (Gneezy and Rustichini, 2004). Non-experimental studies of competition, such as Ehrenberg and Bognanno (1990) and Bandiera et al. (2005), have also focused on the intensive margin of effort rather than on work time.[2]

Moreover, our design uses a field experiment with workers performing a real job that is valuable to the employer (DellaVigna et al., 2016) to incorporate the insights from the recent behavioral economics literature suggesting that individuals might be motivated to stay longer and work harder than their minimum or contracted hours when asked to do so by their employer because of behavioral motivations (for surveys, see Gneezy et al., 2011 and Cooper and Kagel, 2016). Failing to include these motivations in the counterfactual without a bonus could overstate the real-world value of tournament pay, particularly if financial incentives crowd out non-financial impulses (Deci, 1971; Gneezy et al., 2011). Therefore, our field experiment includes several design components meant to trigger those impulses (discussed in Section 2.2). These efforts appear successful in the data: over 58% of workers in the fixed (no bonus) payment scheme worked longer than 11 minutes (Figure 1). Nevertheless, effort is costly to our workers: only 7% worked the full 40 minutes and 30% left within 7 seconds of the minimum time.

Our main finding is that work time is substantially higher among workers competing for a tournament prize bonus of $30 to the worker with the highest output in their pre-defined 4-person work group. In that treatment, 55% worked the full 40 minutes and only 15% worked less than 11 minutes. Workers in the $30 bonus group also invested more effort per minute of work than those who were not offered the chance to compete for a bonus. However, this intensive margin response had a small impact on overall performance relative to the extensive margin

[2] Outside of the literature on competition, a few recent laboratory experiments have used work time as an outcome (Bracha et al., 2015; Linardi and McConnell, 2011; Abeler et al., 2011).

response. Paying a bonus increased total employer costs per work group. To recover these costs, we needed total work time to increase by at least 30%. The average induced increase in effort and performance was greater than 80%. Therefore, our costs per unit of input (work time, effort) or output (task performance) were significantly lower (over 30%) when we paid a bonus.

We focus on tournaments in this paper because of their importance in the workplace, particularly for attaining high-status jobs through promotions, which Lazear (2018) notes "almost always require relative rankings" (p. 202). Within organizations, relative comparisons are often used in employees' formal performance reviews that determine their promotions, salary increases, and layoffs. A prominent example is the practice known of "stacked ranking" or "forced ranking" in which evaluators must conform to a predetermined structure for the overall score distribution, such as shares of employees in highest or lowest categories.[3] Workers may feel further pressure from outside their organizations to invest in work effort; this pressure may come from external competition in labor markets or product markets.

Incentive pay based on individual performance, such as a piece rate, is more common in jobs where individual output is easily measured and closely related to effort, such as manufacturing. Even though we are primarily interested in elite competitive jobs, because we can measure individual effort and output in our setting, we also consider an auxiliary treatment in which all workers are paid a bonus based on their individual performance, with a piece rate value set to match the average employer cost per unit of output in the tournament scheme. We find that paying that piece rate significantly increased time worked and overall effort and reduced costs of effort or output relative to the treatment with no performance-based incentives. In fact, these outcomes were not different from the ones in the $30 tournament. We also explored the effects of awarding a small ($15) prize to the tournament winner; we found significantly higher effort compared to no bonus, but no significant difference in overall effort or costs relative to the higher-stakes tournament or piece rate.

[3] The increased popularity of personnel practices involving relative comparisons with fixed proportions of employees promoted or retained has been attributed to Jack Welch's "rank and yank" approach at General Electric in the 1980s. In recent years, opposition to these schemes has become more prominent, because of their effects on employee morale (see, e.g., Backstone, 2019, on Facebook's system) and possible contribution to sex discrimination (see, e.g., Greenfield and Green, 2017, for reporting on lawsuits against Microsoft, Goldman Sachs and Uber; also see a legal blog on the topic at https://www.law360.com/articles/778682/when-employee-ranking-systems-become-a-legal-liability). Some companies have moved away from strict forced ranking, but it is unclear that the replacement practices are free from any relative comparisons or that they will persist.

Thus, our results clearly show a strong causal relationship between performance-based incentive pay and work hours. This implies that the competitive nature of many high-status and high-pay careers, in which workers compete against their colleagues for bonuses and promotions, is likely to be part of the explanation for their long work hours. Workers in these jobs are not typically paid on an hourly basis or for fixed time shifts. Rather, they have flexibility about how long to work, and can voluntarily provide working hours well beyond the standard workweek in order improve their performance rank. Since workplaces typically hire workers of similar ability levels, even the highest ability workers will need to invest long hours to distinguish themselves from (or not fall behind) their equally capable colleagues.

Furthermore, our results indicate that competitive compensation schemes can be cost-effective and therefore profitable for firms. If they generalize, the induced increase in output that results from the longer working hours (even if only a subset of workers is actively vying for a promotion or bonus) is enough to justify the costs of providing performance-based incentive pay. The profitability of competitive pay is the direct result of the labor supply effects we estimate and can explain its widespread prevalence.

These findings therefore contribute to the literature exploring the reasons for long work hours in elite careers. Prior studies have emphasized explanations such as production technology (Goldin, 2014) or worker signaling of their ability or commitment (Landers et al., 1996, 1997). To focus on the role of tournament incentives in generating long work hours, a phenomenon that has not previously been examined, our study therefore eliminates these alterative explanations by design.

The theory we consider differs from the model in Goldin (2014), which posits a convex production technology, in that competition can be profitable even with a linear technology or diminishing returns to effort. While there are reasons to expect an increasing marginal product of labor at low work hours (from fixed hiring costs or transition costs into tasks for workers), long hours must be accompanied by long working days. Therefore, it is also natural to expect diminishing returns to take effect at some point because of fatigue.[4] The models also differ starkly in their predictions about the future of work hours. Goldin (2014) expresses optimism about the potential for innovations in production technology to reduce the convexity (for

[4] See Hsiang et al. (2019) and citations therein for examples of declining medical care quality later in the day.

example by increasing the substitutability between individual workers) and thereby increase workplace flexibility and reduce work hours. However, if one source of long work hours is worker competition, there is no such expectation.

Our model resembles that in Goldin (2014) in that the additional work hours must be productive in terms of increasing output for the firm. That distinguishes it from the signaling theory in Landers et al. (1996, 1997), in which the value to the firm from the long hours is the information that it provides about workers' types rather than the additional output. Despite this key difference, our model resembles that in Landers et al. (1996, 1997) in that the primary motivator for workers to supply long hours is the desire to succeed in promotion competitions (such as making partner at a law firm, getting tenure in academia or serving in top management at a company). In fact, the central hypothesis of this paper – that incentives from workplace competition increase work hours – can be found in Waldman (1997)'s comment on Landers et al. (1997): competition is proposed as an alternative to signaling as the source of the positive association between work hours and pay.

The idea that competitive incentives increase work hours is also related to Bell and Freeman's (2001) hypothesis that Americans work longer hours than Germans do because greater US wage inequality makes them more concerned about gaining promotions and advancing in the earnings distribution. In that framework, wage inequality increases the value of winning the workplace tournament and therefore the intensity of competition.

Our overall findings for all workers have implications for understanding the gender pay gap. Because women tend to face tighter time constraints from caregiving and home obligations, they may be less willing (or able) to match their male coworkers' work hours and advance professionally. Several recent studies have identified long working hours as an impediment to women's career progress (Bertrand, Goldin, Katz, 2010; Flabbi and Moro, 2012; Gicheva, 2013; Cortes and Pan, 2016, 2017; Mas and Pallais, 2017; Wasserman, 2018). By increasing the number of hours required to advance in elite professional occupations, competition may therefore be contributing indirectly to gender pay gaps by lowering women's representation in those professions (Blau and Kahn, 2000; Gicheva, 2013) and especially in their highest ranks (Bertrand and Hallock 2001; Matsa and Miller, 2011; Kunze and Miller, 2017). The role of workplace competition has previously been examined as a contributor to gender pay gaps (see

Niederle 2016 for a for a comprehensive summary), but not through the channel of extended work hours.

Finally, motivated by the literature finding gender differences in initial entry decisions into tournaments, as well as in performance in certain types of tournaments (see Niederle, 2016), we also examine possible gender differences in the labor supply response to tournament pay. Our design ensures that all workers are available for more than the maximum time, yet men worked significantly longer and harder than women did in our main $30 tournament. There were no significant gender differences in effort under fixed or piece rate pay or in the $15 tournament. To the extent that the gender difference in effort and persistence we find in the $30 tournament also applies to high-stakes workplace competitions more generally, it suggests a further channel through which workplace competition deters women's progress in elite occupations and their ascension to the top ranks of the earnings distribution.

2. Theory and Design of the Field Experiment

The major advantage of conducting a controlled field experiment to study the effects of competition is avoiding the fundamental difficulty with observational data that competitive pay is not randomly distributed across jobs or workers. We do this by having workers perform the same job in the same environment under alternative treatments that vary only in the opportunity and rules for earning a bonus payment. We also control the assignment of workers to payment scheme treatments and offer workers no choice of scheme. In fact, workers assigned to a bonus scheme are informed about it only after they sit down to start their work session. That means that bonus group assignment could not possibly affect a person's decision of whether to accept the job or to show up for work. Our controlled setting also enables us to eliminate alternative sources of long work time proposed in the prior literature by using a non-convex (linear) production technology and a single-shot RA position (with no signaling value).

These aspects of our research design help us isolate and measure the effects of competition on extensive (time) and intensive (effort per unit time) margins of effort supplied by workers. The remainder of this section describes and motivates our other design choices about the work environment, compensation schemes and work task and discusses theoretical predictions.

2.1 Main Treatments: Tournament versus Fixed Payment

In devising our tournament treatment, we start with the requirements that (1) the payment scheme includes financial rewards based on relative performance, (2) performance is a function of effort, along both extensive and intensive margins, and (3) workers are given freedom in setting their effort levels and work time. The details of the treatment, called TP30, are as follows:

TP30: All workers who provide at least 10 minutes of work and then complete a questionnaire about the work are paid $25 for their time. In addition, workers in this treatment compete for a Tournament Prize bonus of $30 paid to one winner from each gender-balanced group of four workers (2 men and 2 women). We set a value of $30 for the prize because it is the smallest round (multiple of 10) dollar amount larger than the promised pay of $25. Competitors in each tournament perform the task simultaneously (receiving their training together and starting at the same time) but independently (on separate tablets) in a small room. The winner of the tournament is the person with the highest total output (defined in Section 2.2). In the event of a tie, the winner is chosen randomly from among those with the highest output. No one except the winner is paid a bonus. Workers are told about the bonus prize after being told they will receive the promised $25 if they work for at least 10 minutes (rather than the hour for which they were hired) as part of the short initial training at the start of the work session. During this training, the task, its purpose, and the value of the work to the employer are discussed (see details below). Workers can work for a maximum of 40 minutes, at which point the program automatically shifts to the questionnaire.

A key feature of the design is that all workers are explicitly told that they are "free to leave" after they finish working and complete the questionnaire. Even though the winner of the tournament bonus can only be determined after the last worker has finished, there is no reason for other workers to stay until the end of the session, because all workers are paid via PayPal within two days of the work session. The option to leave the job site after they stop working, rather than needing to remain for the rest of the hour, substantially increases the opportunity cost to workers of expending effort on the extensive margin.

We estimate the impact of competition on labor supply by comparing outcomes in TP30 to those in a treatment under which workers are given the same options about how long to stay, but not provided with any explicit monetary incentive to work past 10 minutes. This is our fixed payment (FP) scheme.

FP: In the Fixed Payment treatment, workers perform the same task under the same conditions as those in TP30, except that they are paid the same $25 regardless of how long they work beyond the mandatory 10 minutes. Like TP30 workers, FP workers are kindly asked to stay for as long as they can (up to 40 minutes) and to work as hard as they can, because it will benefit the employer, but no additional payment is offered beyond the promised $25.

2.1.1 Theory Considerations for Tournament Treatment

In the classic Lazear and Rosen (1981) setup, the theoretical prediction is clear that we should expect greater effort in TP30 than in FP. The particulars of our setup differ slightly from the main example in Lazear and Rosen (1981) in that our random "luck" term affects output multiplicatively rather than additively and that we have 4 workers in the room, but these differences will not change the qualitative results.

At a Nash Equilibrium outcome, workers in TP30 take their competitors' strategies as given and adjust their effort levels to the point of equalizing the marginal costs and benefits of effort, or until they hit a binding constraint. The marginal cost of effort is the incremental disutility from engaging in the task and forgoing alternative activities during that time. The marginal benefit of effort is the change in the probability of winning (as a result of increased effort) times the increase in utility from winning.

Our experiment allows effort to vary along both extensive (time worked) and intensive (effort per minute) margins, but the distinction is immaterial to the predictions from the Lazear and Rosen (1981) model. Nevertheless, it is useful to note that the time dimension of the problem, a key focus of this paper, resembles a war of attrition. Because workers are in the same room, they can employ strategies that vary with their coworkers' departure times. If we eliminate the random shocks to productivity and variation in the intensive margin of effort (per minute worked), our setting matches the classic war of attrition considered in Hendricks et al. (1988) for 2 players. The winner is then simply the person who stays longer, and he or she is assumed to quit immediately after the loser does. This makes the payoffs equivalent to a second-price all-pay auction. Without random shocks to productivity, there are sharp discontinuities in returns to effort, as chances of winning go from 0 to 0.5 for the move from a narrow loss to a tie and from 0.5 to 1 from a tie to a narrow win. This drives the pure strategy Nash equilibrium to extreme cases in which both workers either leave within a few seconds or stay until the terminal time.

Mixed strategies that involve workers staying for intermediate amounts of time can smooth away jumps in the returns to effort function and may also be possible in equilibrium. In our setting, the existence of a luck component accomplishes the smoothing and allows for interior solutions. Moreover, workers who are the last to remain in their session may still continue to work as long as working increases their chances of winning and the marginal benefit is greater than the marginal cost.

Our research design limits work time to a maximum of 40 minutes, which is also similar to the finite time horizon in Hendricks et al. (1988). This constraint lowers the level of effort supplied by workers whose optimal unconstrained work time is greater than 40 minutes. However, it can also induce offsetting spillover effects. This happens if some workers whose optimal work time is under 40 minutes in the unconstrained equilibrium experience greater returns to effort with the constraint because now their competitors are prevented from working beyond 40 minutes. Those workers will increase their effort levels as a result, making the impact of the constraint on total work effort supplied in the room theoretically ambiguous.

Spillover effects can also induce workers to change their optimal effort levels in response to the distribution of other workers in the room with them. A worker who expects one or more competitors to quit early (e.g., because of high effort costs) will have a higher marginal benefit of staying longer (and higher total benefit of staying the full time) than an otherwise identical worker who expects to face 3 competitors until the end.

Thus, although the direction is clear, the extent of the labor supply response to tournament pay depends on the distribution of worker types and the equilibrium outcome of the game that is played (which may not be unique) and is therefore clearly an empirical question. Furthermore, as we discuss in the next subsection, if we relax the theoretical setup to incorporate agents who are not purely rational selfish optimizing workers, then even the direction of the effect can be ambiguous.

2.1.2 Design Considerations for the Fixed Payment Treatment

In the standard model, workers expend no costly effort beyond the minimum required in their contracts for payment, which makes the FP treatment trivial and potentially unnecessary for

assessing TP30.[5] We have two main reasons for including the FP comparison. The first is simply to confirm that effort is costly to workers, which we do by checking that not all FP workers stayed for 40 minutes and exerted maximal effort. The fact that 30% of FP workers left within 7 seconds of the minimum time further points to costly effort.

Our second motivation for including this treatment is grounded in recent results from the behavioral economics literature suggesting that workers do, at times, work longer than their contractually mandated hours even without explicit monetary compensation. This happens, for example, when they are intrinsically motivated because of characteristics of the work itself (see Gneezy, Meier, and Rey-Biel, 2011 for a recent survey)[6] or when they are the type of person who always works hard ("boy scouts" in Segal, 2012) or when work relationships include elements of gift-exchange (see Cooper and Kagel, 2016 for a recent survey). Because these factors are likely to operate in the workplace, and because their effects on labor supply may be partially or entirely crowded out by offering monetary incentives, we decided it was important to include an FP comparison group for our measurement of the effects of competition.

For the FP comparison to be meaningful, however, we need to use a field experiment with a real work task that has value to the employer (DellaVigna et al., 2016) rather than a real-effort task in a laboratory experiment.[7] Four elements of the field experiment design contribute, in combination, to inducing FP workers to work longer than the minimum time in response to the request to do so.

First, we ensure that workers are available to stay for longer than 10 minutes, and rule out external time constraints as a source of variation in labor supply, by hiring workers for a full hour and then setting a maximum work time of only 40 minutes.

Second, we use a task in which greater effort (on both the intensive and extensive margins) can credibly be described as beneficial to the employer. Our explanation (described in

[5] If a worker faces a threat (explicit or implicit) of being fired for working less than some number of hours above their contracted work hours, we would clearly consider the necessary hours, rather than the nominally contracted hours, to be the relevant minimum number of hours for that job.

[6] We are focused on costly effort and therefore not interested in capturing intrinsic motivation from a work task that is itself enjoyable to workers. Even if workers find parts of their jobs intrinsically motivating and therefore experience periods of low (or even negative) effort costs, this is unlikely to apply to all necessary parts of the job to a degree that exceeds enjoyment from leisure (Lazear, 2018).

[7] While laboratory experiments can allow for a deeper understanding of the operating forces, our concern here is that subjects in a laboratory would treat the tasks they are asked to perform as games rather than work, and we would not be able to induce workplace social considerations.

detail in Section 2.2) is that the employer needs reliable performance data on the computer program that will be used in future research.[8] Workers are told that the purpose of the job is to learn how well people can perform the task under different conditions and asked to, "*please try your best*" (emphasis in the written script; see Appendix B). This makes it clear to workers that investing high effort is expected and necessary for the employer to achieve the goals for which the worker is hired.

The "testing and benchmarking" task also provides a natural justification for the unusual combination of conditions: hiring workers for a full hour, only requiring that they stay for 10 minutes, but then asking them to stay as long as they can. We explain to workers that, although the employer asked them to be available for the full hour, she thought that it "might be too taxing to do this task for so long." In fact, workers are told that figuring out how long individuals can perform the task is one of the reasons why they were hired. Therefore, the instructions state that while the employer "would like you to stay for as long as you can, in order to get paid you only need to perform the task for at least 10 minutes and answer the questionnaire about the task." Not only are workers asked multiple times to try hard and to work as long as possible, but the purpose is also explained to them – it is in order to improve the quality of data received from the testing. Thus, workers are told that staying longer is beneficial to the employer, but it is not necessary for receiving the $25 wage.

Third, we surprise workers favorably about their working conditions (and for some, their compensation). This is intended to trigger positive feelings towards the employer. Telling workers that they only need to work 10 minutes in order to be paid the full amount promised for one hour provides an unexpected "gift" to workers from the employer and may inspire reciprocity motives. These motives should be enhanced by the reason given for the shortened work time, which is the employer's concern for the workers' wellbeing.

Fourth, the employer shows respect and appreciation for the workers' effort. She does this by kindly asking workers to "*please stay as long as you can*" and "*please try your best*" (emphasis in both cases in the written script; see Appendix B). She also does it by explaining to them how their output will be used in future research. These efforts should enhance reciprocity, feelings of duty, and intrinsic motivation, and direct those impulses to be expressed through increased labor supply.

[8] A variant of the program was indeed used by one of the authors in subsequent research.

In practice, as discussed in Section 4 below, our efforts were successful at inducing many workers in FP to stay beyond the required 10 minutes, some for significantly longer: over 58% worked longer than 11 minutes and 7% worked the full 40 minutes (Figure 1). We believe that behavioral considerations are the most likely reason for this additional labor supply in FP rather than dynamic considerations about future employment or recommendation letters from the supervising professor. We attempted to exclude those out in our design by hiring workers for a one-time job, as part of a time-sensitive mass-recruiting drive, rather than an ongoing relationship, and by using a rote and unskilled task.[9]

2.2 The Work Task: Benchmarking the Red Square Program

As described above, workers are hired for research assistance positions aimed at helping to test and benchmark a computer program. The program is generically named Red Square.[10] It is a simple "game" in which players earn points by tapping on stationary squares that appear on a tablet computer (see screenshots in Appendix B). We use a computer program so that we can control the task and automatically track how workers engage with the program to create reliable measures of effort and output. The job of testing a new program also makes it natural to collect end-line survey data about the subjective work experience and opinions of the program.

During a work session, the program alternates between "active" and "rest" screens. At the start of each active screen, a stationary red square appears at a random location. The player earns a point if they tap on the square. Once the red square is tapped, it disappears from the screen, and a button appears that allows the player to advance to the next rest screen. If the player does not tap the advance button, the screen automatically advances to the rest screen 10 seconds after the start of the current active screen, whether or not the red square has been tapped. Each rest screen lasts 10 seconds; there is nothing for players to do during this time. Once 10 seconds have elapsed, the rest screen disappears and the next active screen appears. This cycle repeats until the end of testing. The only variation is that, with a probability of 10%, the active screen includes a gold square (in a random location) in addition to the red one. Tapping the gold square earns 5

[9] We are not able to say for certain what workers expected, but it is worth noting that not a single worker contacted the employer requesting a reference letter or additional employment.

[10] Workers are not specifically told the name of the program, but the name is stored on the work tablets, and so can potentially be discovered by them.

points, so workers can earn 6 points on a screen with a gold square by tapping both gold and red squares.

For the duration of the work session, the following items are displayed on the top left of the screen: running tallies of accumulated points earned and time worked and a countdown timer showing the time left on the current screen. After 10 minutes of testing, a "Go to the questionnaire" button appears on the bottom of the screen. The worker then has the option to tap on the button and end their testing session immediately or to continue working and tap on the button at a later time.[11] After 40 minutes of testing, the questionnaire automatically appears on the screen. Workers can therefore spend between 10 and 40 minutes on the work task. There is no time limit for the questionnaire.

Several features of the task are worth highlighting.

First, the task is extremely simple to understand and to perform. There is no scope for outside knowledge to affect performance and workers need only brief training.

Second, the inclusion of rest screens helps prevent even highly motivated workers from straining or over-exerting themselves. This should make it physically possible for all workers to perform the task for extended periods of time.

Third, the enforced waiting from the rest screen also serves the function of making the task quite boring and tedious, which increases the costs of effort. It is crucial for us that workers not find the task intrinsically enjoyable or fun, because we want to study costly labor supply. The choice of stationary squares, always the same size and colors, are similarly intended to reduce enjoyment of the task. The random elements, varying the location of the red square and only offering gold squares on occasion, might make the game slightly more interesting, but they also increase the attention demanded of players, who have to scan each new active screen to find the square or squares. Responses to the questionnaire item of "How much did you enjoy the game?" confirm that workers did not generally enjoy the task. With a scale from 0 (not at all) to 5 (very much), 61.4% answered 0 (29.2%) or 1 (32.2%).[12]

[11] To avoid mistaken termination from accidental taps, workers are asked to confirm their decision to end the work session before they advance to the survey. The exact wording is "Are you sure you want to stop working on the task? You will not be able to return to the task." The options are: "Cancel" or "Continue to questionnaire."

[12] Workers in FP, who had no financial incentive to stay longer than the minimum time, but who often did, actually expressed the least enjoyment: 73.3% answered 0 (33.3%) or 1 (40.0%).

Fourth, workers can vary the amount of effort they exert within the program along both extensive and intensive margins, and their output (number of points earned) increases in direct proportion to their effort. By tapping the "Go to next screen" button on the active screen faster, a worker earns points faster. They still have to spend 10 seconds on the rest screen, but the next active screen arrives sooner and increases the number of points that can be earned per unit of time worked. At one extreme, a worker who never taps "Go to next screen" will be shown 3 active screens per minute, or 120 over 40 minutes. A worker who instead clicks "Go to next screen" after 1 second, on average, will see about 5.5 active screens per minute, or 218 over 40 minutes.

We are therefore able to use the program to construct three measures of effort. The first is simply the extensive margin of time spent working. The second captures total effort on both margins: it is the total number of times a worker taps the "go to next screen" button over the entire 40-minute work session. For the intensive margin, we also use the frequency of taps, but only include minutes in which the worker is working. We chose the "go to next screen" button instead of the red square (though results are unchanged if we use the latter instead, or if we use total points) because the variation in red square taps is driven almost entirely by variation in active screen time and the resulting frequency of red square offers.

Fifth, the "gold square" feature of the program described above introduces a random "luck" component to the output function that maps effort to points earned. This generates randomness in outcomes, similar to the ε term in Lazear and Rosen (1981), as discussed above in Section 2.1.

Finally, the potential for differences in physical (finger speed) or cognitive (alertness) ability to affect performance is very limited relative to the scope for differences in effort. This means it is practically impossible for a high ability worker to "work smarter, not harder" or to "coast" on low effort and still earn a high score; the only way to earn points is to sit in the room, watch the screen, and tap the squares. It also means that low ability workers should feel that they have a chance to win and therefore expend effort.[13] Although ability variation may be quite important in productive output, most high-stakes workplace competitions include an abundance of high-ability contenders. This makes low-effort a risky (and generally unsuccessful) strategy,

[13] Lazear and Rosen (1981) demonstrate that mixed-ability tournaments with private information about ability will generally lead to inefficient levels of effort. Brown (2011) shows empirically that large skill differences among competitors lower average effort.

even for the most talented among them. Thus, it was important to minimize the role of ability in our work environment.

2.3 Alternative Bonus Schemes

In addition to our main comparison between TP30 and FP, we also study two alternative bonus schemes. As in the first two treatments, workers in these treatments are all paid the promised $25 for staying at least 10 minutes. They may also receive an additional bonus payment that is related to their performance of the task.

***PR*:** Because we can measure effort and performance exactly in our setting, piece rate compensation is possible. We use the PR treatment to test if outcomes differ with individual incentives, based on absolute instead of relative performance. We set the price per point to match the actual average amount paid in bonus per point under the TP30 tournament, which was 3⅓ cents per point.[14] As discussed in Lazear and Rosen (1981), it is theoretically unclear *a priori* how labor supply, worker utility or firm profits will compare between tournament pay and piece-rate compensation.

***TP15*:** Our choice of $30 for the main tournament was based on setting a bonus level above the fixed payment amount. We have no reason to expect that it will be optimal for employers. In particular, by limiting the total work time to 40 minutes, we also limit the amount of incremental effort the employer can extract from each worker. This suggests that a lower prize amount might be equally effective, which we test with TP15, in which a $15 prize is paid to the winner of each tournament. Our expectation is that a lower prize value will weakly decrease the total effort supplied at the room level. We have no such prediction at the individual level, because the $15 difference in prize value could have different utility effects on different workers. If some workers who stayed to the end in TP30 decide to quit early, others who left early in TP30 might respond by increasing their effort to become contenders in TP15.

3. Implementation of the Field Experiment

The field study was conducted at a major American research university in the Spring of 2016. A professor at the university sent emails to departmental undergraduate major email lists with the

[14] As we do not know our workers' degree of risk aversion or their beliefs about competitors' strategies, we are unable to derive the piece rate value that is, on average, equivalent to TP30 from their perspective.

job announcement that invited interested applicants to click on a link to an online survey to apply. The email made it clear that multiple RAs were needed and would be hired for the same position. All recruitment material and scripts are in Appendix B. Applicants had several days to complete the online survey, and the work sessions were held in conveniently located library study rooms on campus. Potential workers were provided a link to a secure website where they can apply for the position and provide contact (name, email address, phone number) and background (gender, major, year, GPA) information and list their periods of availability during the workweek. We used availability and gender for work assignment and asked about year, major and GPA for plausibility.

Conditional on availability and gender, applicants were randomly assigned to one-hour work sessions (particular time slots on particular days) in such a way that each session had an equal number of men and women assigned to that session. Applicants were informed by email that they were hired, provided with the date and time of the work session and the location of a central room used for intake, and asked to confirm their employment by clicking a link and completing an online form.

In this email, they were explicitly told that they would be working in groups and asked to therefore arrive a few minutes early to ensure a timely start. Workers could have assumed before this point that they would be working in groups – because of the initial statement about a large number of workers to be hired ASAP and because of the sign-up form offering fixed time slots for availability – but even if they had not done so previously, it would be clear to all workers before they arrived at the work session that other student RAs would be present as well.

To increase attendance at the work sessions, workers who confirmed employment were sent a reminder the day before the assigned session.[15] Slots that opened up because invited workers declined the invitation were reassigned to other applicants to the extent possible. No applicant was ever assigned to more than one work session.[16]

Workers arrived at the work location and checked in with the manager at the central intake room. Workers were allocated into rooms to maximize the number of 4-person, gender balanced rooms. Workers in all rooms tested the same computer game, but different room-

[15] The average share (across sessions) of confirmed workers who showed-up as scheduled was 91% while the median show-up rate was 92%.

[16] One person managed to sign up twice by using a different email address for signup but the same PayPal account. This was discovered after the fact, and we dropped the room from the analysis.

session combinations were assigned to different compensation scheme treatments.[17] We confirmed the balance of workers' characteristics (as reported in the sign-up form) across treatment, for all workers, and separately by gender. Results, in Appendix Table A1, show no significant differences across treatments (from joint F-tests) in most outcomes, but we do fail the balance test at the 10 percent level (in 3 out of 42 tests) for being a third year student in the pooled and male samples and for being a first year in the female sample. Controlling for these variables in our individual level regressions (pooled or by gender) has no effect on the estimated treatment effects.

At the designated session start time, trained graduate student assistants escorted the workers in groups to their assigned rooms to start the work. Workers sat in front of tablets, and the assistant described the game testing task, reading from the predetermined set of instructions, provided in Appendix B. The assistant then answered any questions, made sure the game was loaded and working on each of the tablets, and left the room. The work was conducted unsupervised unless workers encountered problems.[18] Workers were instructed to leave their tablets on the table when they departed. The last worker to leave each room was asked to send a text message to inform the manager that the room was empty so that the tablets could be secured. Workers were all told they would be paid within 2 days.

After each day of sessions, total amounts owed to each worker were computed, and workers were paid as promised via PayPal. In total, the sessions generated data for an analysis sample of 236 workers. This includes 15 gender-balanced 4-person sessions in FP, TP15 and TP30 and 14 sessions in PR.[19]

4. Effects of Tournament Pay on Labor Supply and Employer Costs

This section discusses our main results from the control (FP) and main (TP30) treatments. The outcomes of interest include three measures of labor supply: duration of work, total

[17] Worker that were assigned to rooms that were not full and balanced were all assigned to the FP treatment and their data are excluded from the analysis.

[18] We dropped all disrupted sessions, 6 in total. The disruptions occurred in early sessions because some workers started the program early and then tried to restart it later. After we become aware of this problem, we instructed the assistant to ensure that all programs were running properly and that all workers started at the same time, which eliminated further problems.

[19] There are only 14 piece rate rooms because one of the 15 sessions included a worker who had previously been hired under a different e-mail address (see footnote 9). We detected the issue only after the work sessions were complete and were therefore unable to add another session.

(unconditional) effort, and intensity of effort (conditional on working). The first two measures are novel to the literature on tournaments as they incorporate the extensive margin of labor supply. The third measure captures the intensive margin and is more standard in the existing literature. We also examine the effects of competition on employer costs in this section. Results from auxiliary treatments and estimation of heterogeneous effects by gender are discussed in later sections.

4.1 Effects on Time Worked

Figure 1 presents histograms of the distributions of work time across individual RAs in the FP and TP30 treatments. It shows two notable features of the FP treatment, discussed in Section 2.1. First, the distribution of work time clearly demonstrates that some workers were induced to work longer than the minimum time with no direct financial incentives to do so. About 58% of the FP group worked longer than 11 minutes and the average time worked was 16.2 minutes. This indicates that our design efforts were successful at getting workers to treat the program testing RA job as they would another job and to therefore be willing to work longer in response to a (justifiable) request from their employer.[20] Second, the fact that only small fraction of FP workers stayed the full 40 minutes confirms that effort was indeed costly to our workers.

Comparing the FP and TP30 distributions in Figure 1 reveals our main result: offering a tournament prize induced people to work much longer. Workers assigned to TP30 worked substantially longer than those not offered a bonus. The median person in TP30 worked for the maximum time of 40 minutes and less than 15% worked for under 11 minutes. This shows that financial incentives based on relative performance increased work time well beyond the effects of behavioral considerations operating in FP.[21]

[20] The presence of behavioral considerations hypothesized in Section 2.1 are confirmed in some workers' statements of the reasons for staying the amount of time they did, such as feeling a moral obligation because they had been hired for a one-hour job (mentioned by 17.9% of workers who stayed for the full time). Other RAs (7.5% of those who stayed less than 40 minutes) apologized for leaving early or provided the excuse that they felt that staying longer would not be helpful to the professor. These statements also suggest that workers perceived that they were expected to stay to help the professor.

[21] It could be that the behavioral factors remained operational in TP30 (and that gift exchange was even enhanced by the opportunity for additional payment). It is also possible that financial incentives crowded out non-financial motivations, but that their effect was much larger, leading to an overall increase in labor supply.

We conducted non-parametric and regression-based tests to assess the statistical significance of the apparent difference between FP and TP30. Although we have data on individual workers, we take a conservative approach to testing and use a room-level unit of analysis. This is because strategic interactions (for TP30) and social norms (for FP and possibly for TP30) within rooms make it likely that outcomes are correlated across workers in the same room. Results are unchanged if we use the individual level instead.

Our regression analysis, presented in column 1 of Table 1, supports the two main findings of Figure 1. The outcome is average time worked per person, and the unit of observation is a room. Because work time is bounded above (at 40) and below (at 10), we estimate Tobit models with upper and lower limits.[22] The regressors of interest are the treatment groups. The omitted category is our main treatment, TP30, so point estimates in the table are all relative to that group.[23] Workers without any monetary incentives for effort work significantly more than the minimum necessary for payment: the constant term + the FP treatment dummy = 16.25 minutes, which is significantly different from 10 ($p < 0.001$).[24] Nevertheless, work time in FP is significantly ($p < 0.01$) and substantially (13.56 minutes per worker) lower than in TP30. Adding performance-pay in the form of a winner takes all tournament increased timed worked by about 80%.[25]

4.2 Effects on Effort

The results in the previous section show that workers stayed longer in TP30, but not that they supplied greater effort or worked more intensively. We consider these outcomes next.

[22] Results are unchanged if we use OLS models that ignore the bounds. Because the room level limits are only binding if all workers in the room are at the minimum or maximum time, we also confirmed that results are unchanged if we use an individual level analysis (with standard errors clustered at the room level) that applies the bounds at the individual worker level (the level at which they are imposed in the experiment). The coefficients in the individual-level Tobit models suggest a larger increase in work time between FP and TP30, making our main room-level estimates a conservative measure of the impact of competition on work time.

[23] To keep the models and results consistent throughout the paper, results for all treatments are in Tables 1 and 2, though discussion of auxiliary treatments (PR and TP15) only starts in Section 5.

[24] A Wilcoxson signed rank test on room-level data from FP rejects the hypothesis that time stayed is equal to 11 ($p = 0.01$).

[25] The corresponding non-parametric comparison of these work time distributions shows the same results. The distribution of room level average time worked in TP30 first-order stochastically dominates the distribution in FP; the Kolmogorov-Smirnov test yields $p < 0.001$. The opposite test (that FP dominates TP) is not significant, with $p = 1$ in the Kolmogorov-Smirnov test.

The observable action that we use to signify effort is tapping the button to advance to the rest screen and our effort measure is the number of taps. If we instead use the number of total taps (also including taps on red and gold squares) as an effort measure, or if we only add taps on red squares (with no random component) to our effort measure, the results are qualitatively and quantitatively unchanged. Our main regression models have a room-level unit of observation (as in the previous section). We aggregate effort across workers and over the full session time and use two measures of effort. The first captures average effort per worker, where effort is the number of times a worker taps the "go to next screen" button in the session. The second isolates the intensive margin of effort supplied per worker by only including workers who are still formally "on the job" in the denominator. The measure is created by dividing the total effort supplied by workers in the room during the session by the total time worked in the session.

To graphically explore how effort evolved over time, we also calculated room-level effort measures separately for each of the 40 minutes in the work session and then averaged these values across rooms by treatment category. Figure 2 displays the first measure, which is average effort per worker provided in each minute. In this measure, workers naturally contribute zero effort in minutes after they end their session. By contrast, those workers are excluded from the second effort measure, displayed in Figure 3, average effort per worker per minute among workers who are still working at that minute. We calculated that value for each minute using only individuals who worked for the full minute: we then divided their total effort during that minute by their number. The effort measures in Figure 2 and Figure 3 are the same for the first 10 minutes when all four workers are working, but this changes after workers start to leave.

Figures 2 and 3 clearly echo the two main patterns in the prior section: workers in the FP treatment continued to supply effort after the 10 minutes necessary for payment and effort was significantly higher in TP30.

It is also apparent that allowing the extensive margin to vary in our field experiment was an important design feature, as that margin is by far the more economically significant one. Differences in the intensive effort measure are small, especially after the mandatory 10 minutes. Part of this might be because the sample of workers used to compute the intensive margin in Figure 3 is changing over time as workers who supply less extensive margin effort are dropped when they stop working. If these workers also supply less effort on the intensive margin, the sample of current workers becomes more favorably selected over time. Such selection effect

would be larger in FP than TP30 because more workers leave early in FP. We can eliminate this source of bias by focusing on the first 10 minutes of the session (in Figure 2 or 3): for those minutes, the graphs do suggest higher effort in TP30 than FP.

A concern about focusing on the first 10 minutes is that effort might be changing during the session if workers get better at the task with practice or if they slow down when they get tired. Figure 3 shows no evidence of those dynamics: average effort (within treatment) was fairly constant after the first few minutes. A very similar picture appears if we remove selection effects by restricting attention to those who stayed for the full work time (giving us a balanced panel on a fixed set of workers). This lends supports to the value of examining effort in the first 10 minutes and further suggests that our production technology is indeed linear. Learning (if it happens) is limited to the first 3 minutes and fatigue did not hamper performance within 40 minutes of work.

We quantify the magnitudes of these effects using room-level regressions, and report results in columns 2, 3, and 4 of Table 1. Starting with the total effort measure, we find a substantial effect of competition on effort. Total room effort is 88% higher in TP30 than in FP (column 2).[26] Conditional on working, the intensive margin of effort is also significantly greater in TP30 than in FP ($p < 0.05$ in column 3 for the entire session; $p < 0.10$ in column 4 for the first 10 minutes),[27] but the magnitude of the increase is fairly small. Relative to TP30, there is a mere 4% reduction in mean effort intensity, either over the entire session (column 3) or within the first 10 minutes (column 4).

In unreported estimates, we also considered a different type of effort measure based on mistakes, defined as red or gold squares that are displayed but go "untapped." These mistakes turned out to be very rare in our data. Out of 30,585 red squares shown on screens across all treatments, only 45 (0.15%) were not tapped. The fraction of red squares missed, the number of red squares, and an indicator for at least one red square missed in a room are all uncorrelated with the treatment. Workers were slightly more likely to miss tapping a gold square, but still the

[26] Using non-parametric tests, we find that the total effort level in TP30 first-order stochastically dominates that in FP (with Kolmogorov-Smirnov tests yielding $p < 0.001$), while total effort in FP does not dominate that in TP30.

[27] Using non-parametric tests for both intensive margin measures, we find that effort in TP30 first-order stochastically dominates the level in FP (with Kolmogorov-Smirnov tests yielding $p = 0.014$ for the full session and $p = 0.091$ for the first 10 minutes) while effort in FP does not dominate that in TP30 (both Kolmogorov-Smirnov tests yield $p = 1$).

number is small. Out of 3,001 gold squares shown on screens in all treatments, 37 (1.2%) were not tapped. As with the red squares, the fraction of gold squares missed, the number missed, and an indicator for at least one being missed in a room are all uncorrelated with the treatment. Because these mistakes reflect effort invested on the intensive margin, the finding that they are uncorrelated with treatment supports the limited intensive margin response we found with our main effort measures. In light of these results, we focus on measures that include the extensive margin (time worked and total effort) in what follows.

4.3 Effects on Employer Costs

The previous two sections document large and significant increases in work time and effort in TP30 relative to FP, but the additional labor was not free. Rather, the increased effort was induced by paying a $30 prize to the worker with the most points in their room, which increased labor costs by 30%. This raises the question of whether paying for performance was worthwhile to the employer.

Table 2 presents the answer to this question using several alternative measures of employer costs of effort or output. Column 1 examines labor costs per work minute, using the average amount paid per minute worked in the room. We find that TP30 significantly decreased labor costs per minute. Without the bonus, we paid $2.18 more for each minute worked, which corresponds to 46% higher costs per minute. In column 2, we use payment per red tap because the objective of the task was to earn points and red taps capture points earned irrespective of luck. We find the same result: TP30 was significantly cheaper than FP. We paid $0.11 more (or almost 50% more relative to the TP30 mean) per red tap in FP. In column 3, we return to our main effort measure from Section 4.2, the number of times workers tapped the “go to next screen” button, and measure costs per tap. Again, the results are the same: we paid $0.121 (52%) more for effort in FP. Results are also unchanged if we measure costs per point earned in the game (where each red square equals 1 point and each gold square equals 5 points; column 4), per tap of either a red or gold square (column 5), or for any action that could measure effort (i.e., tapping a red square, gold square or the “go to next screen” button; column 6). In each case, we paid significantly (at least 50%) more in FP than when we provided workers with performance-based incentives. (Scaled to the higher average effort costs incurred in FP, the tournament prize reduced costs by over one-third.)

These measures show an economic value to employers of offering tournament incentives, which is the reduction in the cost of extracting effort from workers. Another benefit that applies in this setting may not generalize to other workplaces, but is worth noting: the improved quality of the "testing" data produced. Recall that one of our stated goals to workers was learning how long individuals could (physically and mentally) perform the red-square task. The answer to that question differs dramatically if we use results from FP or TP30. In FP, only 6.67% workers stayed the whole 40 minutes and 15% stayed 30 minutes or longer. In TP30, however, 55% stayed the full 40 minutes and 61.7% stayed 30 minutes or longer. Based on TP30, it appears that most workers could work for at least 40 minutes. If we had only run FP, we would have instead concluded that at most 10% of workers could do that. Although this particular outcome is specific to the RA tasked used in this study, it is worth noting that work quality (from the view of what the employer values) in our setting was also higher in TP30 than FP.

5. Auxiliary Treatments

In light of the large effects of TP30 on labor supply and employer costs, it is natural to ask if such a large prize was necessary to induce effort from workers. This is particularly relevant in our setting where work time was capped at 40 minutes and many TP30 workers worked that long. We assess this using the TP15 treatment that differs from our main TP30 treatment only in awarding a $15 rather than $30 prize to the winner.

Because output and effort can be measured accurately in our task, it was also feasible to offer individual piece rate incentives that can be effective at motivating workers to supply more effort. We separately designed a PR treatment with a payment per point set to 3⅓ cents to match the amount that was paid in the TP30. As discussed in Section 2.3, this price is set to match the average employer costs across all workers, but it is not expected to match workers' expected marginal returns to effort functions between the two treatments. Returns to effort are independent across workers in PR but depend on what others are expected to be doing in TP30. The results from these alternative treatments are shown in Figure 4 and Tables 1 and 2.

Figure 4 depicts the CDF of room-level total time worked by treatment. Like Figures 1 and 2, Figure 4 again depicts the longer work time in TP30 relative to FP that was discussed in Section 4.1. Interestingly, we find that work time is distributed nearly identically in TP30 and PR, with 48.2% of workers staying the full 40 minutes in PR. Work time appears somewhat

lower in TP15 compared to the other incentivized treatments, but still substantially longer than in FP, with 41.67% of workers staying the full 40 minutes.

The regression results in Table 1 show the same patterns. The omitted category is TP30. Estimates in column 1 indicate that, while work time (in minutes per worker) was shorter in TP15 and slightly longer in the PR relative to TP30 (coefficients of -3.9 and 0.49, respectively), these differences are statistically insignificant. Table 1 also reports the relevant F-tests between each of the different treatment pairs. These tests indicate longer work times relative to FP in each of the three treatments that includes performance-based incentives but no significant differences among those treatments.[28]

Column 2 reports estimates with the total effort measure that captures both extensive and intensive margins.[29] As in column 1, we find no significant differences in total effort invested across the different treatments with financial incentives. The point estimates suggest lower effort in both TP15 and PR relative to TP30, but the differences are not statistically significant. The F-tests, reported at the bottom of the table, further show that total effort in the auxiliary treatments was significantly higher than in FP and also that TP15 and PR are not significantly different from one another.

The next two columns present results for the intensive margin effort measures: for the whole 40 minutes (in column 3) and for the first 10 minutes (when all workers are present and working, in column 4). For each of these measures, we find that neither auxiliary treatment is different from TP30, or from the other. The one difference that emerges in columns 3 and 4 is for tournaments relative to FP: workers invested significantly more intensive margin effort in TP15 and TP30, but not in PR. Nevertheless, as the main impact of performance-based incentives occurs along the extensive margin of time worked, the pattern is still consistent with higher total effort in PR than in FP (as in column 2) and no significant difference in total effort between PR and the tournaments.

[28] Again, the non-parametric tests deliver the same results. In both the TP15 and PR treatments, work time was significantly longer than in FP. The distribution of time worked in either of these treatments first-order stochastically dominates FP (the Kolmogorov-Smirnov tests yield $p = 0.005$ and $p < 0.001$, respectively), but is not dominated by it (both the Kolmogorov-Smirnov tests yield $p = 1$). However, Kolmogorov-Smirnov tests and Mann-Whitney tests indicate that the TP30, TP15, and PR are not different from one another.

[29] Interested readers can see the graphs for both intensive margin effort measures for all 4 treatments in Appendix Figures A1 and A2.

The fact that both TP15 and PR increased labor supply relative to FP (reported in Table 1) suggests that the employer might have also been able to lower costs with the auxiliary treatments. This is shown in Table 2, which reports the various cost measures described in Section 4.3. Across the various measures of costs per output or effort, the auxiliary treatments are never significantly different from TP30 (i.e., the TP15 and PR coefficients are never significant). However, costs are significantly lower in the new bonus treatments TP15 and PR than they are in FP for every output and effort measure we consider (shown in the F-tests below the coefficients). The auxiliary treatments are also not different from one another (as indicated by the *p*-values for the last F test).

The findings from the auxiliary treatments therefore provide further empirical support for the theoretical prediction that tournaments can lead to longer work time, while also showing that similar results can be achieved with individual incentives (when available to employers). The fact that TP15 and TP30 had similar effects on labor supply and employer costs shows that outcomes may not be overly sensitive to the precise details of the incentive scheme.[30] The fact that tournaments can improve profits even if the employer is not able to solve for (or implement) the optimal prize structure provides additional support for their widespread use in practice.

6. Gender Differences in Labor Supply Across Treatments

In this section, we explore gender differences in labor supply within and across treatments. We focus on the extensive margin of effort, time worked, because that is the main driver of variation in total effort (Section 4.2), but also show that results are unchanged for total effort.[31] Our unit of observation is now an individual, so we account for within-room correlations by clustering standard errors at the room level.

We start by estimating separate effects of competition on the labor supply of male (columns 1 and 2) and female (columns 3 and 4) workers. The results for time worked are depicted in Figure 5. Relative to TP30, we find that both male and female workers provide

[30] Additional comparisons between TP30 and the alternative bonus schemes are presented in Online Appendix C. Section C.1 shows that effort is significantly higher in TP30 than in TP15 if we expand our model to control for "luck" (rate of gold stars per screen), suggesting that higher prizes induce more effort. Section C.2 shows a significantly lower rate of having a single worker in the room at 40 minutes in TP30 and PR, consistent with the spillovers across workers implied by the tournament structure.

[31] We find very few differences when examining our intensive margin measures of average effort per minute worked, overall or in the first 10 minutes. These results can be found in Appendix Table A2.

significantly less effort in FP. The *p*-values, reported at the bottom of the table, indicate that both male and female workers invested more effort in TP15 and in PR relative to FP. This indicates that men and women both respond to financial incentives based on either relative or individual performance metrics.

However, Figure 5 also shows that responses differ by gender. Among the performance-based incentive schemes, men provided the most effort in TP30, followed by PR, and then TP15. Women, on the other hand, provided the most effort in PR, followed by TP15 and TP30 (which appear nearly identical in the figure). We quantify these observations with regression estimates in the first four columns of Table 3. While the ordering is indeed as depicted in Figure 5, the only significant difference in effort provided across the incentivized treatments is for men between the two tournaments: men invest significantly less effort in the lower-stakes TP15 than in TP30. The other differences are statistically insignificant.

Because Figure 5 groups outcomes by gender, differences between men and women are not readily apparent. We therefore show the distributions grouped by treatment in Figure 6. We also report estimated gender differences by treatment group in a regression framework by supplementing the separate regressions for male and female workers in columns 1-4 of Table 3 with results from a pooled sample of men and women in columns 5 and 6. In these pooled models, the treatment dummies capture the differences across treatments (relative to TP30) for men.[32] The coefficients for female interacted with each of the 4 treatment dummies correspond to the gender differences within each treatment.

Figure 6 and the additional regression results indicate that men and women worked the same amount of time in FP and PR (coefficients on Female×FP and Female×PR are small and insignificant). Although the figure suggests gender differences in both tournament treatments, the regressions reveal the only significant difference is in TP30. Specifically, in TP30, women worked significantly less than men did (coefficient of -11.69, $p = 0.044$) and as a result invested less effort (coefficient of -35.47, $p = 0.053$). If the gender difference in work time and effort in TP30 applies to high-stakes competitions more generally, this result implies that women will be underrepresented among tournament winners, and hence less likely to be promoted to high ranking positions, even without external constraints on their work time.

[32] The observant reader will note that the treatment dummies in Column 5 are not identical to the ones in Column 1. The reason is that we use a non-linear Tobit model for work time to account for censoring.

The fact that we find a gender difference in effort in TP30 but not in TP15 is consistent with the prior literature on competition showing that gender differences vary depending on the specific features of the competition, such as the prize amount (Petrie and Segal, 2013) or nature of task (see Niederle, 2016 for a summary). However, because it presents a more complex picture, we further confirmed that the differential between the two tournaments was not the result of some spurious correlation from luck (being worse for women in TP30 or for men in TP15) or worker characteristics to treatments. Our balance checks in Appendix Table A1 indicate that background characteristics and luck are, on the whole, equal across treatments by gender. Nevertheless, we repeated the regressions in columns 5 and 6 of Table 3 adding controls for characteristics and luck (the fraction of screens shown that included a gold square and the square of this value); the analysis confirms the same pattern of gender differences.

To understand the source of the observed differences in effort between TP30 and TP15, we must first remember that each worker's effort level is an equilibrium outcome in which that worker is responding optimally to their beliefs about the strategies of their 3 competitors. The marginal (financial) return to effort is the prize amount multiplied by the marginal increase in the probability of winning the prize from an increase in effort. Increasing the prize amount would always increase the returns to own effort if competitors kept their effort unchanged, but it seems likely that at least some competitors would also find the increased prize attractive and increase their effort. In particular, the value to any given worker of staying the entire time is lowered with each additional coworker who is expected to also stay that long. It would therefore be wrong to think that workers whose effort is unchanged are ignoring financial incentives because their expected financial incentives might not have changed much.[33]

The presence of a gender gap in TP30 and absence of one in TP15 together suggest that responses to tournament incentives are determined differently by gender. Because of the equilibrium effects described above, the pattern we observe can be generated by some utility factor that either (1) attracts men to TP30 more than to TP15 or (2) attracts women more to TP15

[33] Consider a worker whose effort costs make them willing to stay 40 minutes for an expected bonus of $12 or higher. That worker will not compete in TP15 if they expect at least one other worker in their room at 40 minutes, because their expected bonus from staying that long is $7.5. The same worker would be willing to compete for 40 minutes in TP30 against one other person (expected bonus of $15) but not against 2 or 3 others (expected bonus of $10 or $7.5). If they faced one competitor in both TP15 and in TP30, that worker would supply more effort in TP30 than in TP15. However, if competition increased the number of competitors staying to the end from 1 in TP15 to 2 in TP30, the worker would supply the same low effort level in both.

than to TP30. Because we ran a field experiment, we cannot directly measure beliefs and preferences and with their help determine whether it is men or women who are responding differentially to TP30 versus TP15 or pinpoint the precise reasons for their behavior. To identify potential sources, we therefore turn to the strand of the experimental literature on gender and competition that, like this paper, assigns workers to different incentive schemes.[34]

One possibility is gender ratios. Starting with Gneezy, Niederle and Rustichini (2003), the literature documents that, under certain conditions, women respond significantly less to tournament incentives in mixed-gender competitions than men do. Although both TP15 and TP30 start out mixed-gender, financial considerations alone should cause fewer workers to stay to the end in TP15. Women in TP15 can therefore anticipate a higher chance of participating in an all-female contest in TP15 than in TP30. If the equilibrium number of workers staying 40 minutes is 2 in TP15 and 3 in TP30, a woman who stays in TP15 has a one-third chance of facing only female competition, but one who stays in TP30 has no chance. Indeed, one-third of all TP15 sessions ended with female workers engaged in single-sex tournaments, while all TP30 sessions had at least one male worker among the last to leave (Fisher exact test yields $p = 0.042$). Thus, our female workers may have found TP30 relatively less attractive because it involved competition with men and TP15 relatively more attractive because of competition with another woman (Gneezy, Niederle and Rustichini, 2003; Niederle, Segal and Vesterlund, 2013).

It is also possible that male workers derived more utility from competing in TP30 than in TP15, and were willing to increase effort in it, because the higher prize amount made the competition more exciting and salient to them. Iriberri and Rey-Biel (2017) document a large improvement (relative to piece rate) in male, but not female, performance in tournaments where competition is made salient, but not in other tournaments.

Gender differences in beliefs and risk aversion can also produce differential responses to competitive pay. The literature on gender differences in tournament entry, going back to Niederle and Vesterlund (2007), finds these are important contributing factors.[35] In our setting, if women believe that men perform better on the task, this could increase their willingness to

[34] We focus on that strand because workers in our setting face a choice between working under a given incentive scheme or not working and leaving the workplace. They are not given a choice among different incentive schemes (as is done, for example, in Niederle and Vesterlund, 2007, and Dohmen and Falk, 2011).

[35] Recent work by Gillen et al. (forthcoming) even argues those factors explain the entire gender difference in entry.

compete against other women rather than against men. It is in fact a common finding that gender differences in competition arise primarily in tasks for which men are stereotypically expected to outperform women (for a summary, see Niederle, 2016). Additionally, differences in the number of active competitors can also affect the variance in payoffs between TP15 and TP30, which could produce gender differences in effort if men and women differ in their degree of risk aversion. For example, if women are more risk averse than men, they might be willing to stay 40 minutes for a 50% chance of $15 but not for a 25% chance of $30, while men are willing to stay for both. Because mixed-sex competition, high salience, stereotypically male tasks, and high variance in payoffs are all part and parcel of high-stakes competition in the workplace, the gender difference found in TP30 could extend to elite occupations if its source is any of these underlying mechanisms.

7. Conclusions

Long work hours are pervasive in high-pay, high-status jobs. While this feature has previously been attributed to production technology and sorting into professions, we examine whether it could be explained by the competitive incentives common to those workplaces. That is, could on the job competition drive long work hours? The findings from our study suggest the answer is yes.

We use a field experiment at a real job to isolate the effects of competitive pay on employees' work time and effort. Our comparison fixed payment scheme offered workers no financial incentives for additional effort beyond the mandatory work time, yet a majority of workers stayed longer after being asked to do so, suggesting some nonfinancial, behavioral motives are present. Work time and effort were significantly higher in our main treatment with a large tournament prize and lowered costs for the employer by more than a third. Auxiliary treatments with a piece rate bonus or a small tournament prize also induced higher worker effort and lower employer costs relative to fixed wages.

When tournament incentives are cost-effective, it is natural that firms will offer them, and competitive jobs will entail long work hours. Therefore, our findings highlight a fundamental challenge for policy efforts aimed at reducing work hours in order to enhance overall worker wellbeing or improve gender equality in elite occupations. These policies may be rendered ineffective with pay incentives that encourage workers to voluntarily supply additional labor and

circumvent formal hours restrictions.[36] Moreover, policies effective at reducing hours may entail substantial costs to employers and some employees.

Our field experiment also revealed a significant gender difference in labor supply, but only in the high-prize tournament. The fact that men worked longer in that tournament hints at a second channel through which high-stakes workplace competition can contribute to gender gaps in labor markets. The first is based on our overall result that indicates that competition can increase the work hours needed to succeed in elite professional careers. This disproportionately deters women, who traditionally have more binding time constraints from greater household obligations. The second channel is through differential effects of high-stakes competition on labor supply, even absent any differences in time constraints.

[36]See, e.g., Fargen and Rosen (2013) on under-reporting of duty hours among medical residents.

References

Abeler, J., Falk, A., Goette, L., & Huffman, D. (2011). Reference points and effort provision. American Economic Review, 470-492.

Backstone, Nick (January 13, 2019). "Facebook's job evaluations are so ruthless that 'meeting most' expectations could lead to getting fired, former employees say," Business Insider. https://www.businessinsider.com/former-facebook-employees-describe-brutal-job-performance-reviews-2019-1

Bandiera, Oriana, Iwan Barankay, and Imran Rasul (2005). Social preferences and the response to incentives: Evidence from personnel data. Quarterly Journal of Economics 120.3: 917-962.

Bell, L. A., & Freeman, R. B. (2001). The incentive for working hard: explaining hours worked differences in the US and Germany. Labour Economics, 8(2), 181-202.

Bertrand, M., Goldin, C., & Katz, L. F. (2010). Dynamics of the gender gap for young professionals in the financial and corporate sectors. American Economic Journal: Applied Economics, 2(3), 228-255.

Bertrand, Marianne, and Kevin F. Hallock. "The gender gap in top corporate jobs." ILR Review 55.1 (2001): 3-21.

Blau, Francine, and Lawrence Kahn (2000). Gender Differences in Pay. Journal of Economic Perspectives 14(4): 75–99.

Bracha, A., Gneezy, U., & Loewenstein, G. (2015). Relative pay and labor supply. Journal of Labor Economics, 33(2), 297-315.

Brown, J. (2011). Quitters never win: The (adverse) incentive effects of competing with superstars. Journal of Political Economy, 119(5), 982-1013.

Bull, Clive, Andrew Schotter & Keith Weigelt (1987). "Tournaments and Piece Rates: An Experimental Study. Journal of Political Economy 95: 1-33.

Cooper, David J. and John H. Kagel .2016. Other-Regarding Preferences: A Selective Survey of Experimental Results. In John H. Kagel, Alvin E. Roth (Eds.), The Handbook of Experimental Economics, Volume 2, Chapter 4, 217–289. Princeton: Princeton University Press.

Cortes, Patricia and Jessica Pan (2016). Prevalence of Long Hours and Women's Occupational Choices. Working Paper.

Cortes, Patricia and Jessica Pan (2017). When Time Binds: Returns to Working Long Hours and the Gender Wage Gap Among the Highly Skilled. Working Paper.

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. Journal of Personality and Social Psychology, 18(1), 105.

DellaVigna, Stefano, John A. List, Ulrike Malmendier, and Gautam Rao (2016). Estimating social preferences and gift exchange at work. National Bureau of Economic Research Working Paper w22043.

Dohmen, Thomas, and Armin Falk (2011). Performance Pay and Multidimensional Sorting: Productivity, Preferences, and Gender. American Economic Review, 101(2): 556-90.

Ehrenberg, Ronald and Michael Bognanno (1990). "Do Tournaments Have Incentive Effects?" Journal of Political Economy 98(6): 1307-1324.

Fargen, Kyle and Charles Rosen (2013). Are Duty Hour Regulations Promoting a Culture of Dishonesty Among Resident Physicians? Journal of Graduate Medical Education, 5(4): 553-555.

Flabbi, Luca and Andrea Moro (2012). The effect of job flexibility on female labor market outcomes: Estimates from a search and bargaining model. Journal of Econometrics 168 (2012) 81–95.

Freeman, Richard B., and Alexander M. Gelber (2010). Prize Structure and Information in Tournaments: Experimental Evidence. American Economic Journal: Applied Economics, 2(1): 149-64.

Gicheva, Dora. "Working long hours and early career outcomes in the high-end labor market." Journal of Labor Economics 31.4 (2013): 785-824.

Gillen, Ben, Erik Snowberg and Leeat Yariv (forthcoming). "Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study," Journal of Political Economy.

Hendricks, Ken, Andrew Weiss, and Charles Wilson. (1988). "The War of Attrition in Continuous Time with Complete Information." International Economic Review 29(4): 663-80.

Hsiang, E.Y., Mehta, S.J., Small, D.S., Rareshide, C.A., Snider, C.K., Day, S.C. and Patel, M.S. (2019). Association of Primary Care Clinic Appointment Time With Clinician Ordering and Patient Completion of Breast and Colorectal Cancer Screening. JAMA Network Open, 2(5): e193403-e193403.

Gneezy, Uri, Stephan Meier, and Pedro Rey-Biel. (2011). "When and Why Incentives (Don't) Work to Modify Behavior." *Journal of Economic Perspectives* 25(4):191-210.

Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in Competitive Environments: Gender Differences. The Quarterly Journal of Economics, 118(3), 1049-1074.

Gneezy, Uri, and Aldo Rustichini (2004). Gender and competition at a young age. American Economic Review 94.2: 377-381.

Goldin C. (2014). A Grand Gender Convergence: Its Last Chapter. American Economic Review. 104(4):1091-1119.

Greenfield, Rebecca and Jeff Green. November 8, 2017. "Uber's Employee Ratings Put Women at a Disadvantage, Suit Says." Bloomberg News. https://www.bloomberg.com/news/articles/2017-11-08/uber-s-employee-ratings-put-women-at-a-disadvantage-suit-says

Iriberri, Nagore, and Pedro Rey-Biel (2017). Stereotypes are only a threat when beliefs are reinforced: On the sensitivity of gender differences in performance under competition to information provision. Journal of Economic Behavior & Organization, 135: 99-111.

Kunze, Astrid, and Amalia R. Miller (2017). Women Helping Women? Evidence from Private Sector Data on Workplace Hierarchies. Review of Economics and Statistics, 99(5): 769-775.

Landers, R. M., Rebitzer, J. B., & Taylor, L. J.. (1996). Rat Race Redux: Adverse Selection in the Determination of Work Hours in Law Firms. American Economic Review, 86(3), 329–348.

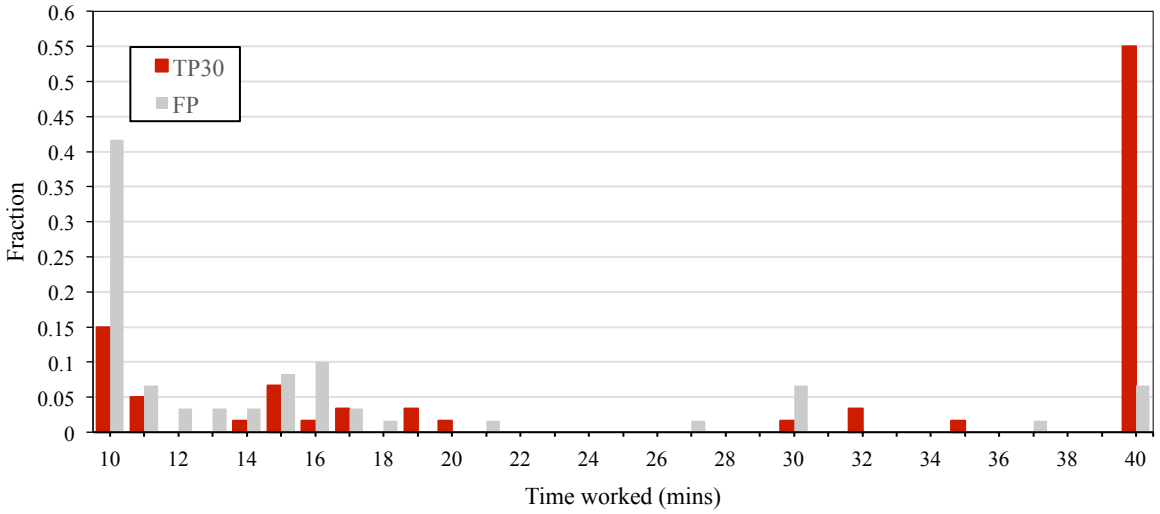
Landers, R. M., Rebitzer, J. B., & Taylor, L. J. (1997). Work norms and professional labor markets. In Gender and family issues in the workplace, ed. F. D. Blau and R. G. Ehrenberg, 166–202. New York: Russell Sage Foundation.

Lazear, Edward & Sherwin Rosen (1981). "Rank-Order Tournaments as Optimum Labor Contracts." Journal of Political Economy 89: 841-864.

Lazear, Edward P. (2018). "Compensation and Incentives in the Workplace." Journal of Economic Perspectives, 32 (3): 195-214.

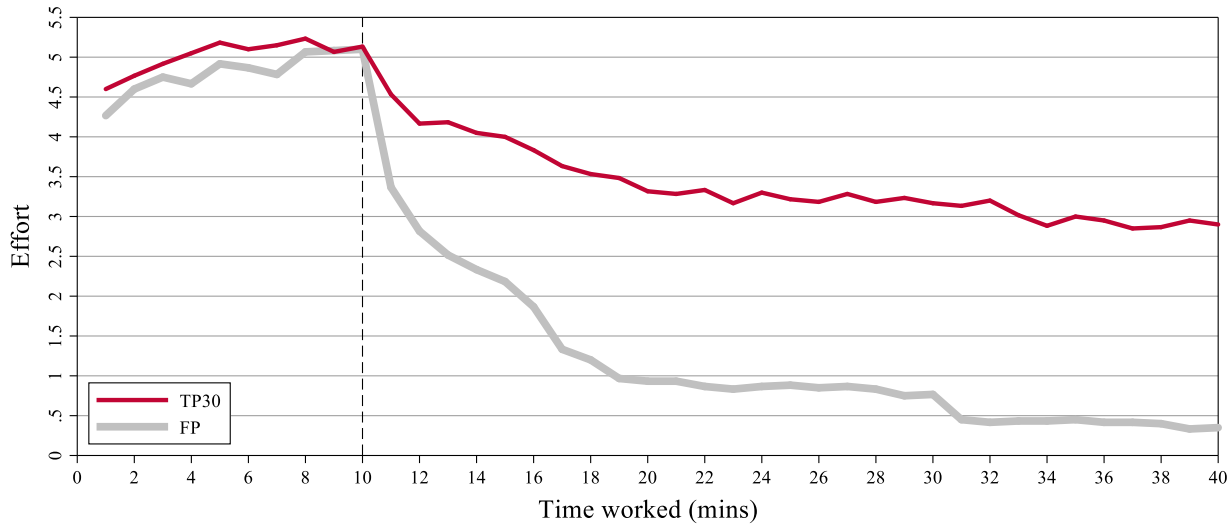
Linardi, S., & McConnell, M. A. (2011). No excuses for good behavior: Volunteering and the social environment. Journal of Public Economics, 95(5), 445-454.
Mas, A., & Pallais, A. (2017). Valuing alternative work arrangements. American Economic Review, 107(12), 3722-59.
Matsa, David A. and Amalia R. Miller (2011). Chipping Away at the Glass Ceiling: Gender Spillovers in Corporate Leadership. American Economic Review, 101(3): 635–39.
Niederle, Muriel. (2016). Gender. In Handbook of Experimental Economics, second edition, Eds. John Kagel and Alvin E. Roth, Chapter 7, 481-553. Princeton: Princeton University Press.
Niederle, M., Segal, C., & Vesterlund, L. (2013). How costly is diversity? Affirmative action in light of gender differences in competitiveness. Management Science, 59(1), 1-16.
Niederle, M., & Vesterlund, L. (2007). Do Women Shy Away From Competition? Do Men Compete Too Much?. The Quarterly Journal of Economics, 122(3), 1067-1101.
Petrie, Ragan and Carmit Segal (2017). Gender Differences in Competitiveness: The Role of Prizes. Working Paper.
Segal, Carmit (2012). Working when no one is watching: Motivation, test scores, and economic success. *Management Science*, *58*(8), pp.1438-1457.
Waldman, M. (1997). Commentary on Chapter 6. In Gender and family issues in the workplace, ed. F. D. Blau and R. G. Ehrenberg, 166–202. New York: Russell Sage Foundation.
Wasserman, Melanie (2018). Hours Constraints, Occupational Choice, and Gender: Evidence from Medical Residents. UCLA Working Paper.

Figure 1: Histogram of Time Worked



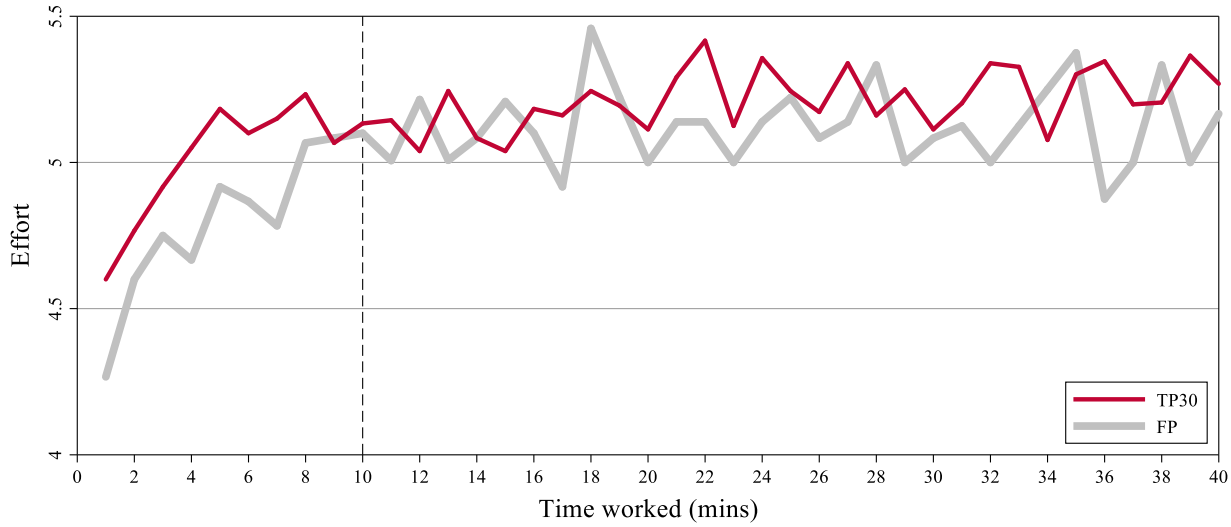
Notes: Distribution of work time across individual workers in TP30 ($30 tournament prize) and FP (fixed payment) treatments. Work time was constrained to lie between 10 and 40 minutes.

Figure 2: Mean Room Effort per Minute



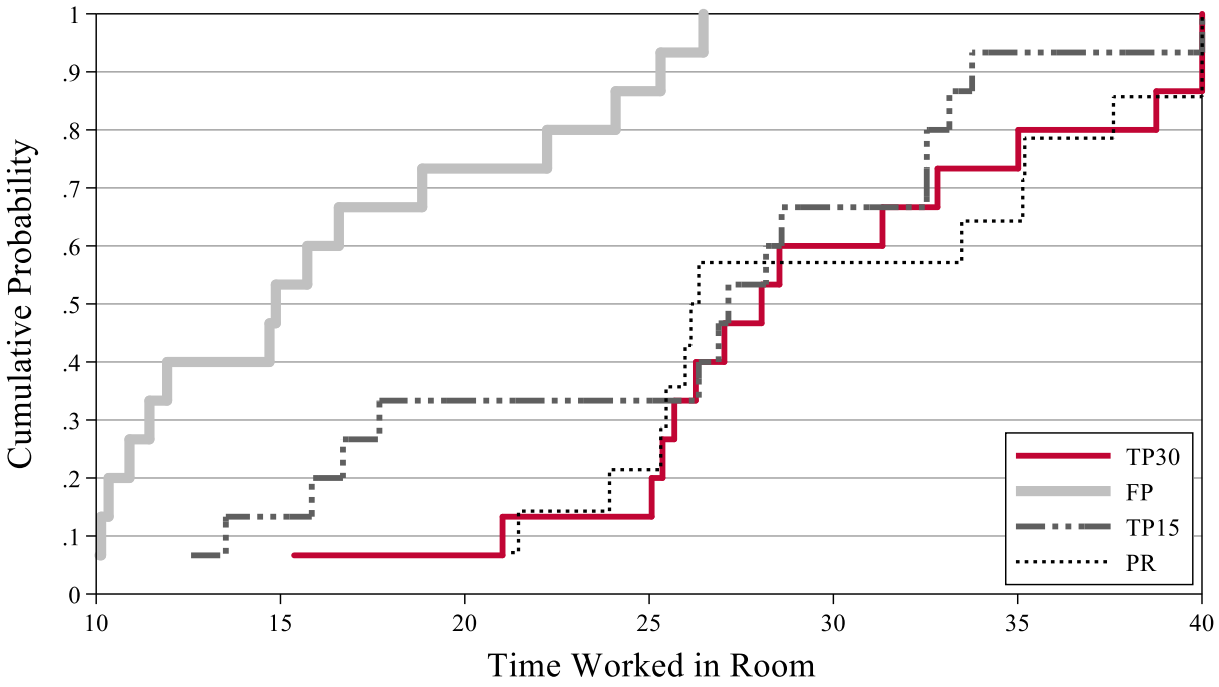
Notes: Average room-level effort per minute of work session in TP30 ($30 tournament prize) and FP (fixed payment) treatments.

Figure 3: Mean Room Effort Conditional on Working per Minute



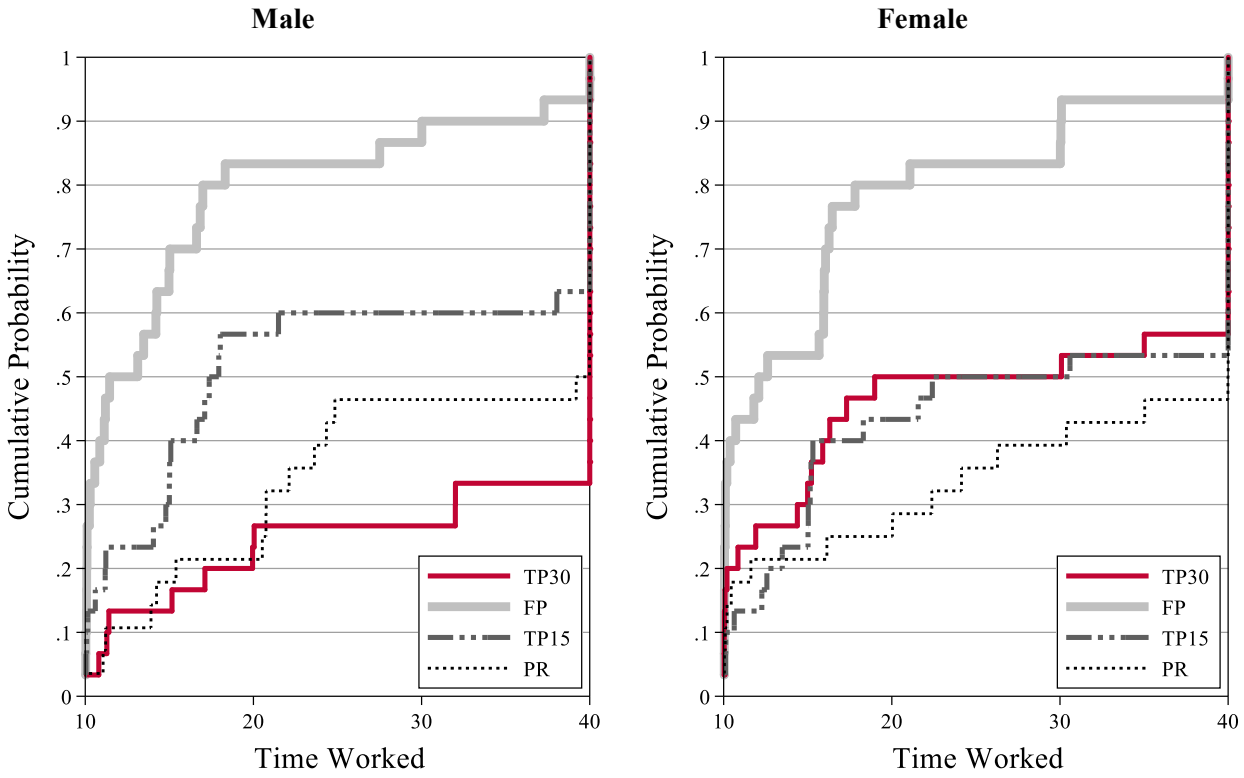
Notes: Average room-level effort per minute of work session, among workers still on the job for that minute, in TP30 ($30 tournament prize) and FP (fixed payment) treatments.

Figure 4: CDF of Average Time Worked in Room by Treatment



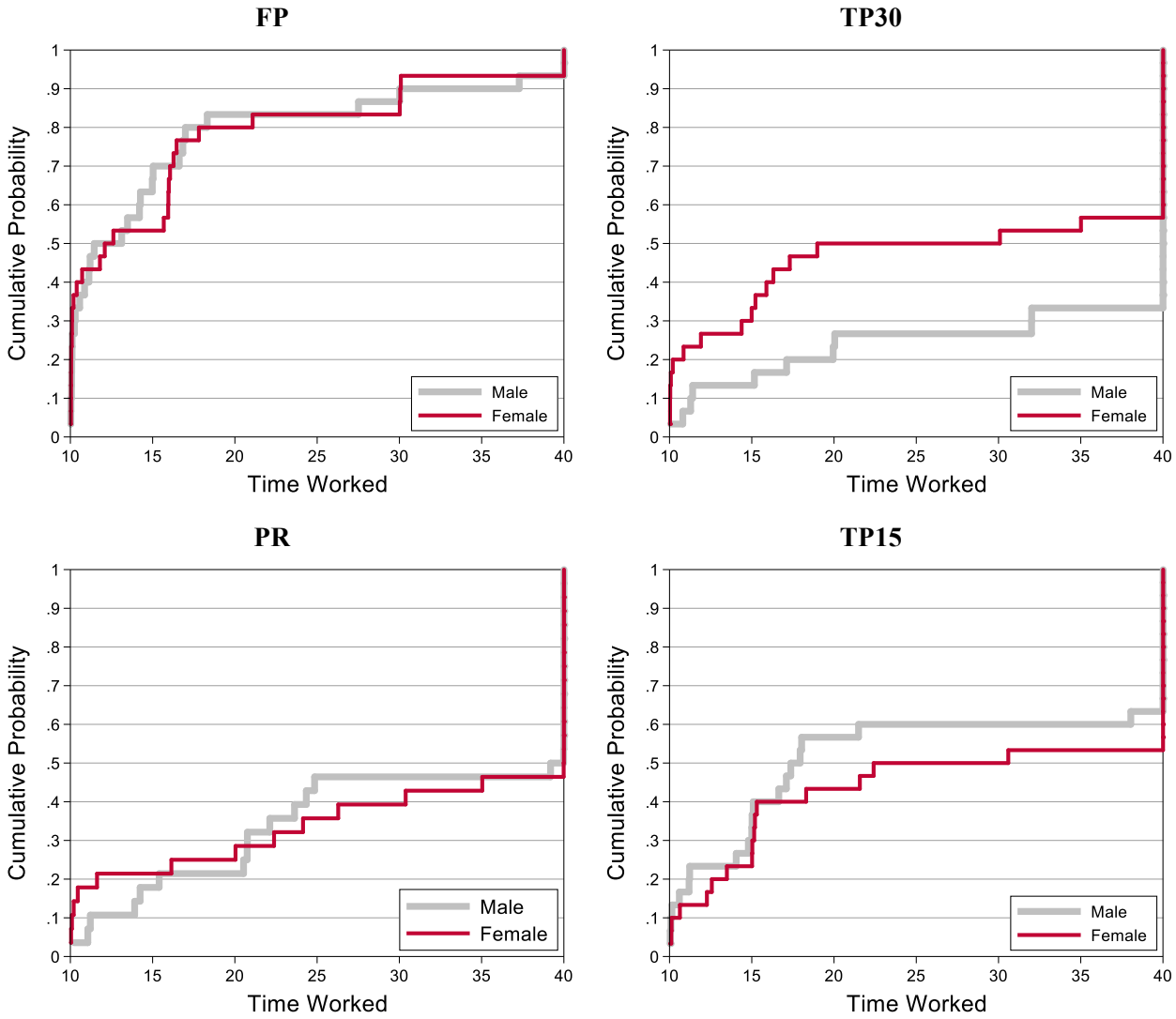
Notes: CDF of room-level average work time per session in TP30 ($30 tournament prize), FP (fixed payment), TP15 ($15 tournament prize) and PR (piece rate) treatments.

Figure 5: Treatment Differences by Gender



Notes: CDF of individual work time per session in TP30 ($30 tournament prize), FP (fixed payment), TP15 ($15 tournament prize) and PR (piece rate) treatments, separately by gender.

Figure 6: Gender Differences by Treatment



Notes: CDF of individual work time per session by gender, separately for TP30 ($30 tournament prize), FP (fixed payment), PR (piece rate), and TP15 ($15 tournament prize) treatments.

Table 1: Effects of Compensation Scheme on Room Level Labor Supply

	(1)	(2)	(3)	(4)
	Time Worked	Total Effort in the Session	Effort per Worker-Minute Conditional on Working	Effort per Worker-Minute in the First 10 Minutes
		Main Treatments		
FP	-13.562***	-70.867***	-0.214**	-0.210*
	[2.494]	[12.687]	[0.086]	[0.121]
		Auxiliary Treatments		
TP15	-3.921	-19.017	-0.025	0.033
	[3.011]	[15.412]	[0.087]	[0.110]
PR	0.494	-1.587	-0.112	-0.109
	[2.859]	[13.511]	[0.075]	[0.126]
Constant	29.807***	151.033***	5.123***	5.020***
	[2.023]	[9.881]	[0.058]	[0.090]
Prob>F: TP15 = FP	< 0.001	< 0.001	0.0415	0.0218
Prob>F: PR = FP	< 0.001	< 0.001	0.201	0.405
Prob>F: TP15 = PR	0.15	0.25	0.281	0.196
Observations	59	59	59	59
R^2		0.384	0.130	0.196

Notes: Observations at the room level. Tobit model is estimated for time worked and linear regressions for effort measures. Robust standard errors in brackets. *** p < 0.01, ** p < 0.05, * p < 0.1

Table 2: Effects of Compensation Scheme on Labor Costs

	(1)	(2)	(3)	(4)	(5)	(6)
	Pay for Work Minute	Pay for Tapping a Red Square	Pay for Effort (Tapping "Go to Rest Screen")	Pay for Point	Pay for Tapping a Red or Gold Square	Pay for Any Tap (Red, gold, or "Go to Rest Screen")
			Main Treatments			
FP	2.181***	0.114***	0.121***	0.081***	0.105***	0.056***
	[0.678]	[0.035]	[0.037]	[0.026]	[0.033]	[0.017]
			Auxiliary Treatments			
TP15	0.364	0.018	0.020	0.011	0.016	0.009
	[0.626]	[0.033]	[0.034]	[0.023]	[0.030]	[0.016]
PR	-0.179	-0.006	-0.006	-0.007	-0.007	-0.003
	[0.401]	[0.022]	[0.023]	[0.016]	[0.020]	[0.011]
Constant	4.709***	0.229***	0.232***	0.156***	0.210***	0.110***
	[0.344]	[0.019]	[0.020]	[0.014]	[0.018]	[0.009]
Prob>F: TP15 = FP	0.024	0.020	0.018	0.017	0.019	0.019
Prob>F: PR = FP	< 0.001	< 0.001	< 0.0011	< 0.001	< 0.001	< 0.001
Prob>F: TP15 = PR	0.338	0.398	0.397	0.353	0.384	0.390
Observations	59	59	59	59	59	59
R^2	0.244	0.245	0.250	0.248	0.246	0.248

Notes: Observations at the room level. Pay is total pay, counting fixed payments plus any bonuses. Robust standard errors in brackets. *** p < 0.01, ** p < 0.05, * p < 0.1

Table 3: Labor Supply Results by Gender

	(1)	(2)	(3)	(4)	(5)	(6)
	Male		Female		All	
	Time Worked	Total Effort	Time Worked	Total Effort	Time Worked	Total Effort
FP	-25.577***	-89.833***	-14.000***	-51.900***	-25.769***	-89.833***
	[4.693]	[14.795]	[5.007]	[17.262]	[4.695]	[14.827]
TP15	-14.120**	-43.533**	1.434	5.500	-14.238**	-43.533**
	[6.099]	[20.085]	[6.028]	[20.192]	[6.217]	[20.128]
PR	-6.985	-19.302	5.366	16.129	-7.068	-19.302
	[5.656]	[16.057]	[6.134]	[19.109]	[5.712]	[16.091]
Female x TP30					-11.694**	-35.467*
					[5.775]	[17.940]
Female x FP					0.164	2.467
					[2.018]	[9.328]
Female x TP15					3.967	13.567
					[5.484]	[19.461]
Female x PR					0.714	-0.036
					[4.829]	[14.688]
Constant	42.259***	168.767***	30.872***	133.300***	42.464***	168.767***
	[4.452]	[11.195]	[4.743]	[14.999]	[4.367]	[11.219]
Prob>F: TP15 = FP	0.032	0.020	0.001	<0.001		
Prob>F: PR = FP	<0.001	<0.001	<0.001	<0.001		
Prob>F: TP15 = PR	0.237	0.237	0.500	0.557		
Observations	118	118	118	118	236	236
R^2		0.242		0.153		0.199

Notes: Observations at the worker level. Tobit models are estimated for time worked and linear regressions for total effort (over the session). Robust standard errors clustered at the room level in brackets. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Appendix A: Additional Figures and Tables

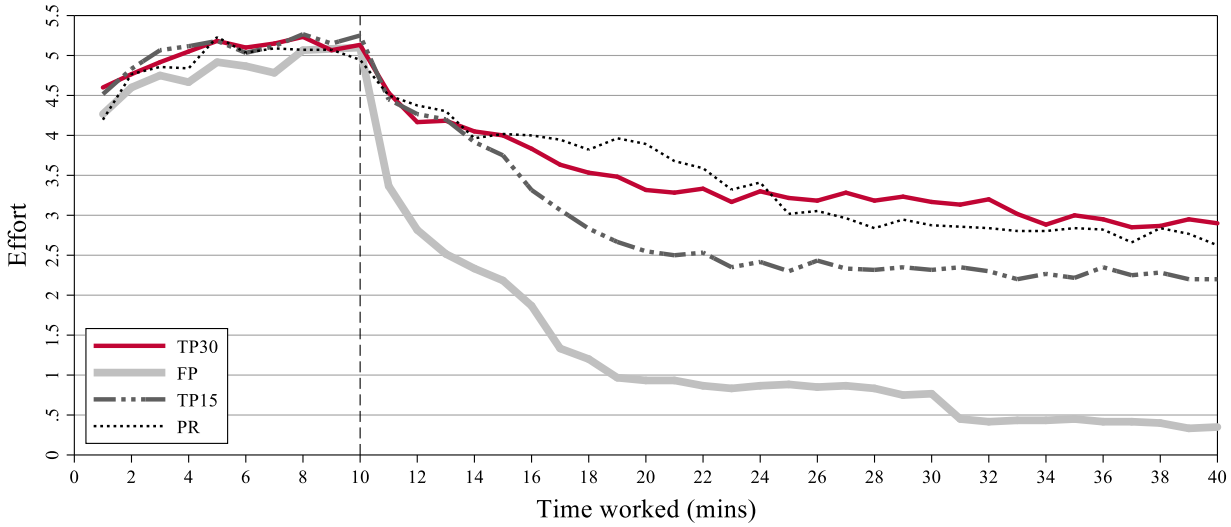


Figure A1: Total Room Effort per Minute – All Treatments

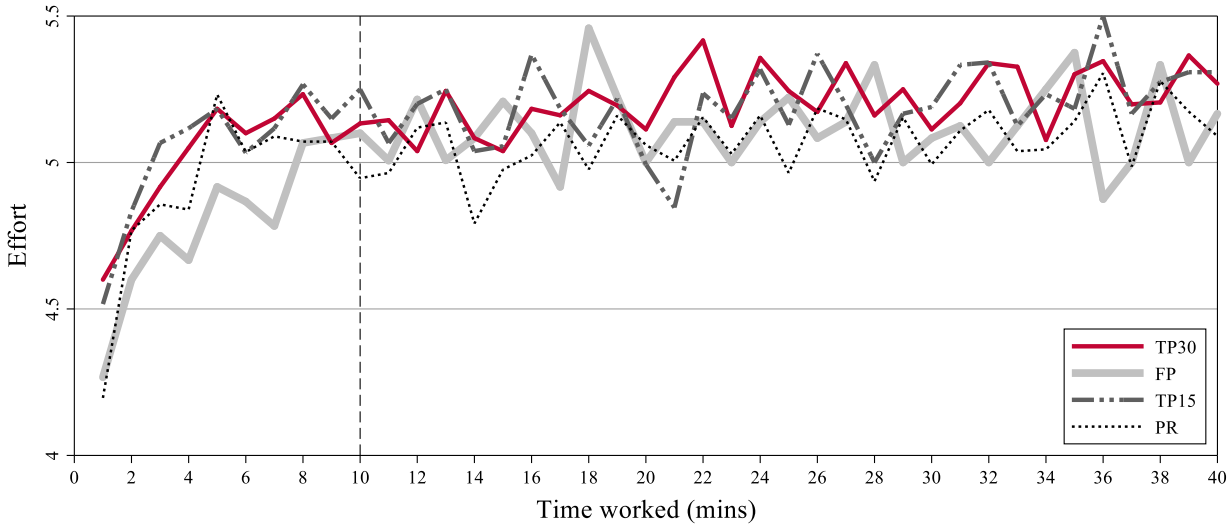


Figure A2: Mean Room Effort Conditional on Working per Minute – All Treatments

Appendix Table A1: Balance Tests

	(1)	(2)	(3)
Sample	All	Male	Female
GPA	0.166	0.518	0.245
GPA-Squared	0.172	0.508	0.254
Economics or Commerce Major	0.503	0.298	0.988
Engineering or Science Major	0.931	0.703	0.992
Other Major	0.722	0.820	0.845
White	0.940	0.789	0.612
Asian	0.948	0.828	0.601
Other Race	0.591	0.882	0.693
First Year	0.109	0.173	0.0208
Second Year	0.179	0.412	0.279
Third Year	0.00539	0.0514	0.215
Fourth Year	0.662	0.403	0.473
Fraction of Screens with a Gold Square	0.630	0.374	0.384
Fraction with Gold Square Squared	0.769	0.460	0.343

Notes: Table reports *p*-values from F-tests on the joint significance of the treatment indicator variables (FP, TP15 and PR) following separate regressions on each of the outcomes listed in the rows. Column 1 uses all workers; column 2 is limited to men; and column 3 is limited to women.

Appendix Table A2: Intensive Margin Effort Measures by Gender

	(1)	(2)	(3)	(4)	(5)	(6)
	Male		Female		All	
	Mean Effort	Effort in the First 10 Minutes	Mean Effort	Effort in the First 10 Minutes	Mean Effort	Effort in the First 10 Minutes
FP	-0.227	-0.310*	-0.168**	-0.110	-0.227	-0.310*
	[0.138]	[0.180]	[0.081]	[0.133]	[0.139]	[0.181]
TP15	-0.021	0.010	-0.004	0.057	-0.021	0.010
	[0.104]	[0.110]	[0.085]	[0.144]	[0.104]	[0.110]
PR	-0.063	-0.228	-0.099*	0.009	-0.063	-0.228
	[0.115]	[0.187]	[0.059]	[0.112]	[0.116]	[0.188]
Female x TP30					0.020	-0.080
					[0.074]	[0.099]
Female x FP					0.080	0.120
					[0.126]	[0.184]
Female x TP15					0.038	-0.033
					[0.075]	[0.093]
Female x PR					-0.017	0.157
					[0.079]	[0.154]
Constant	5.093***	5.060***	5.113***	4.980***	5.093***	5.060***
	[0.090]	[0.099]	[0.049]	[0.103]	[0.090]	[0.100]
Prob>F: TP15 = FP	0.085	0.047	0.088	0.207	0.085	0.047
Prob>F: PR = FP	0.201	0.709	0.344	0.211	0.202	0.709
Prob>F: TP15 = PR	0.644	0.156	0.217	0.665	0.645	0.157
Observations	118	118	118	118	236	236
R^2	0.046	0.055	0.050	0.017	0.049	0.042

Notes: Observations at the worker level. Mean effort is total effort in the session divided by minutes worked. Robust standard errors clustered at the room level in brackets. *** p < 0.01, ** p < 0.05, * p < 0.1

Appendix B: Field Experiment Materials

B.1 Recruitment email

Subject: Hiring for Paid Short-Term Research Assistance Positions in the Upcoming Weeks

Dear students,

I am looking to hire a large number of students ASAP to help me test and benchmark a computer program that will be used in research to measure economic preferences. I would like to learn about user experience and performance under different parameters of the program.

Students invited to work on the project will be paid $25 for testing the program and responding to a questionnaire about the work. All work will be done using tablet computers that I will provide during the work session. Plan on being available for an hour. Payments will be made via PayPal within 2 days after the work session.

Work on the project will begin on Monday, February 15, 2016 and will take place over the next couple of weeks.

If you are interested, please provide information about yourself and your availability as soon as possible through this online form: <survey link>

The deadline to complete the online form is Wednesday, February 10, 2016 @ 12

Thanks!
Professor X

B.2 Sign-up Survey

Sign-up Page for Professor X's RA Positions

Thank you for your interest in Professor X's program testing and benchmarking short-term research assistant positions. To apply for a position, complete the following form to provide information about yourself and your availability.

Students invited to work on the project will be paid $25 for testing the program and responding to a questionnaire about the work. All work will be done using tablet computers that will be provided during the work session. Plan on being available for an hour. Payments will be made via PayPal within 2 days after the work session.

If you have any questions, please contact Professor X by email at professorx@abc.edu with the subject line "Program testing RA Position Question."

Personal Information

1. First name
2. Last name
3. Paypal email (We will pay assistants via PayPal. Please enter the email you use for PayPal below.)
4. Contact email (Please enter the email you want us to use to contact you here.)
5. Phone number
6. Sex (male/female)
7. Birth date (mm/dd/yyyy)

Academic Information

1. University ID number
2. Year in school
3. Major
4. GPA at university
5. Is there any other information you want to provide about yourself? If so, use the space below
6. Availability – please indicate all time slots when you could be available to work (M-F, 9:30-10:30am, 11am-12pm, 12:30-1:30pm, 2-3pm, 3:30-4:30pm)
7. If you are selected for this task, the data that you enter, as well as de-identified data about your work performance and compensation, may be used for research by the professor and coauthors. Check this box to indicate that you consent to have your data used in these ways. <click box>

B.3 Invitation to work session email

Subject: Program Testing RA Position

Dear <Name>,

You have been selected by Professor X as one of her RAs. Please come to Library Room <number> on <date> at <time> to perform the work**.**

You will be working in groups, so please make sure that you arrive a few minutes earlier so that we can start on time and you do not delay your colleagues.

Please make sure that your Paypal account is linked to this e-mail address. Otherwise, I will not be able to pay you.

You were assigned this day and time based on your availability. Please fill out the following form by 5pm on Thursday, February 11 to confirm you will be able to work at the assigned date and time.

<survey link>

If we do not hear from you by then, we will assume you cannot work and will find a replacement.

Thanks,
Research assistant

B.4 Work session instructions (read out loud by research assistant)

Script for Program Testing and Benchmarking RA Position Intro

<As students arrive>
Thanks for coming. Please sit at a tablet computer that you will use for the job.

<When everyone is seated>
I am XX, a PhD student in the department, working with Professor X on this project. I have some information and instructions that Professor X wants me to provide you about the position and the task before you get started.

The main task that you are going to complete is to play a simple game on a tablet computer. In the game, your job is to earn as many points as you can by clicking squares that appear on the screen. Once you log in to the game, it will alternate between "active screens" that have squares and "rest screens" with no squares. For this group, the active screen will last for at most 10 seconds. If you don't click a square during that time, the program will go to a rest screen. If you click on a red square before the 10 seconds are up, you earn 1 point and you are given the option to "fast forward" to the start of the next rest screen. Also, with a 10% probability (so on average in 1 out of every 10 screens), you will also see a gold square. You get 5 points for clicking the gold square. You can also click the red square in the same active screen.

As was explained in the recruitment e-mail, Professor X wants to use this program in her research. For that she needs to get a benchmark of the distribution of response times and how people score under different parameter conditions and time lengths. **So, please try your best.**

She also wants to learn about how the program functions and the user experience. Therefore, once you are done with the task, the program will ask you to complete a short questionnaire about your work experience today.

While Professor X asked you all to be available for a full hour, she thought that it might be too taxing to do this task for so long. This is another thing that she is trying to figure out. Therefore, while she would like you to stay for as long as you can, in order to get paid you only need to perform the task for at least 10 minutes and answer the questionnaire about the task. Once you have finished the questionnaire, you will be free to leave. This is how it will work. After 10 minutes of testing, a button will appear in the upper right-hand corner of the screen. You can click that button to end the program testing and move to the questionnaire. If you choose not to move to the questionnaire, after 40 minutes, the program will automatically end the program testing and direct you to the questionnaire. So, you will need to work for at least 10 minutes and are free to stay and work up to 40 minutes. **Please stay as long as you can.**

<For Tournament treatment:> To provide you with additional incentives to try your best, Professor X decided to pay a bonus in your group. The person that gets the most points, will receive [$30, $15] in addition to the promised $25, for a total of [$55, $40]. In case of a tie, the winner will be chosen randomly from among those who have the most points. This will be done with a roll of a die.

<For Piece Rate treatment:> To provide you with additional incentives to try your best, Professor X decided to pay a bonus in your group. You will be paid additional money for every point that you earn, with an exchange rate of 10 cents for every 3 points. You will be paid a bonus of 3 and one third cents for every single point you earn. This bonus payment is in addition to the promised $25, so your total payment will be $25 plus whatever bonus you earn.

ANY QUESTIONS?
<If there are no questions move to program>

Please login to the program. **Please use the same PayPal e-mail address you gave Professor X for the registration form.** She will not be able to pay you if you use a different email address.

Once you login you will see the parameters of the program for this group. Please wait on this page while I complete the instructions. Does everyone see the information about parameters? I will go around to make sure that there are no problems.

In case you encounter problems with the program please text Research assistant at the number written on this sheet of paper <Indicate paper>.

Once you start the program, you must not exit or minimize the program or turn off the tablet for any reason. If you do that, you may not be able to complete the task and you will not get paid.

Once you are done with the questionnaire, and submit your answers, you are free to leave. Please leave your tablet on the table. If you are the last person in the room to leave, please text Research assistant so he can come collect the tablets.

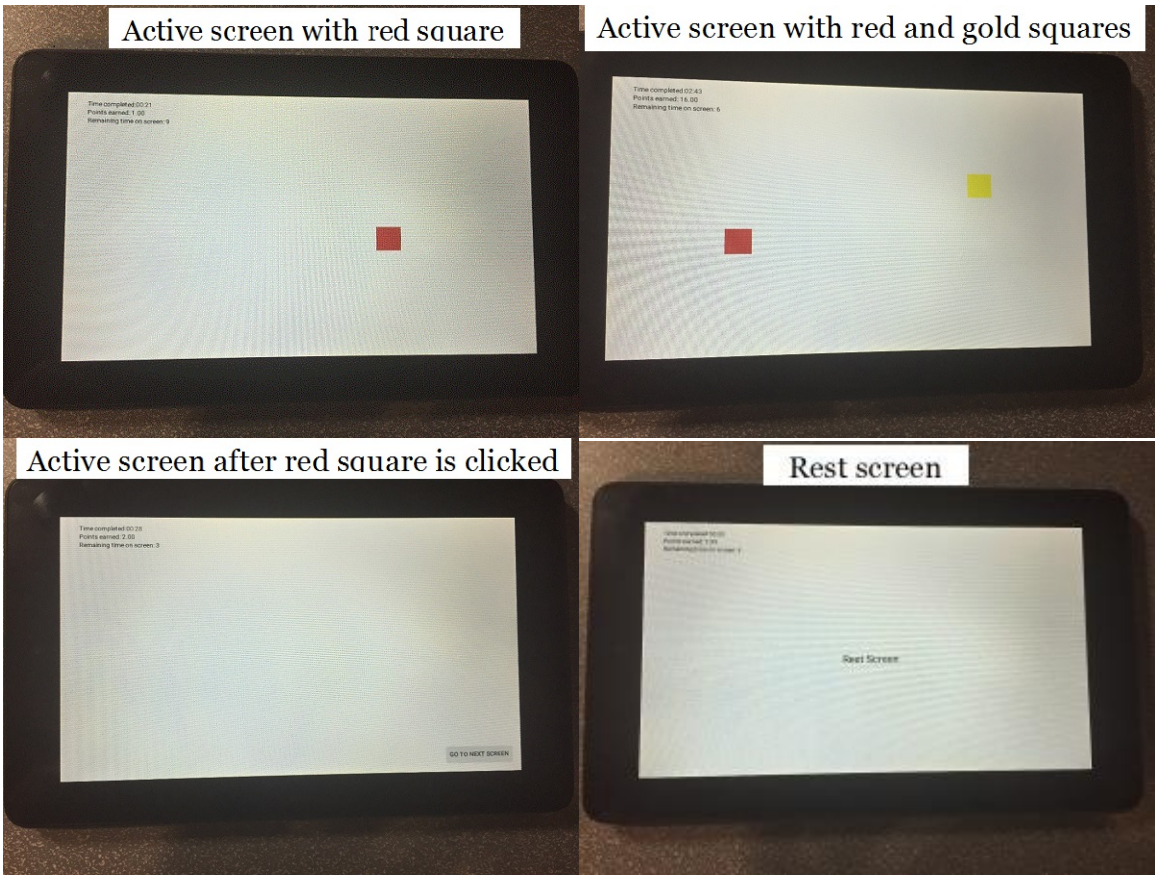
You will receive your payment on your Paypal account. If there are any problems with payments, please contact Professor X.

ANY QUESTIONS?

<RA goes around to check that everyone had the parameter screen up and is ready to start the task>

Please press the "**Start Task"** button to begin.

B.5 Red Square program – screen shots



B.6 Post work session questionnaire

1. How much did you enjoy the game? (0-not at all, ..., 10-very much)
2. How exciting did you find the game? (0-not at all, ..., 10-very much)
3. Did you have enough time for the active screen? (0-not at all, ..., 10-very much)
4. Did you have enough time for the rest screen? (0-not at all, ..., 10-very much)
5. How tiring did you find the game? (0-not at all, ..., 10-very much)
6. Do you have any suggestions for improvements in the game design? (open ended)
7. Why did you decide to leave when you did? (open ended)
8. Would you be willing to come back for another trial under the same conditions? (yes, no)
9. How was the work environment? Mark all that apply. (Friendly, Quiet, Relaxing, Meditative, Boring, Stressful, Painful, Exciting)
10. Do you have any suggestions for improvement in terms of the work environment? (open ended)
11. Did you experience any technical problems (tablet/stands/program)? (open ended)
12. Do you have any suggestions for improvement in terms of technical problems? (open ended)
13. Are you happy with the payment scheme? (open ended)

Appendix C. Detailed Comparison of Main and Auxiliary Treatments

The results in Section 5 show very similar overall effects on labor supply and costs from the alternative bonus schemes. This appendix compares the different treatments in greater detail.

C1. Comparison between TP30 and TP15

It is natural to expect effort to be weakly lower in TP15 relative to TP30. The increase in prize value between TP15 and TP30 suggests an increase in labor supply. However, as we discussed in the main text, the number of competitors is not fixed. Thus, if multiple workers increase their effort in response to the increase in prize, then the probability of winning for each of them is reduced, and that would have a mitigating effect.[1] Furthermore, if a significant fraction of workers is at a corner (i.e., given the number of actual competitors in their room, they were willing to work even longer than 40 minutes for the $30 bonus), then this would also mitigate the effect of the prize on labor supply.

We do find lower effort (total and time worked) supplied in TP15 relative to TP30, but the estimates are not significant (Table 1, columns 1 and 2). In non-parametric tests, we also find fewer workers worked the full time in TP15 than in TP30 (41.67% versus 55%) and more rooms in which all workers left before 40 minutes (4 versus 2), but again, these differences are statistically insignificant.

In a supplemental analysis reported in Appendix Table C1 below, we find statistically significant differences in work time between TP15 and TP30 when we control for random variation in "luck" across workers in the experiment. The probability of being offered a gold square was 10% for each screen, but the realized share of gold squares offered varied across workers. This variation was not systematic across treatments (see Appendix Table A1), but even random noise would reduce the precision of our main estimates if luck affects labor supply. For

[1] It may be worth noting that the spillover effect from competitors' additional effort in TP30 can also *increase* the returns to effort for some workers and cause them to increase effort more between TP15 and TP30 than they otherwise would have. An example can illustrate this. Consider a room with 2 workers with high effort costs who quit immediately in both TP15 and TP30, a 3rd worker with moderately high effort costs who quits immediately in TP15 but is willing to stay 40 minutes for a ½ chance of winning TP30, and a 4th worker with low effort costs who is willing to stay 40 minutes for a ½ chance of winning TP15 or TP30. Equilibrium in TP15 can have all the workers leaving early, including the one willing to stay 40 minutes (because she wins soon after the 3rd worker leaves). A switch to TP30 could produce an equilibrium in which effort is still low for the 2 high cost workers, but both the 3rd and 4th workers stay the full 40 minutes. In this case, both the 3rd and 4th workers increase effort *more* in TP30 when the equilibrium changes than they would have if competitor effort had stayed the same.

tournaments, it is plausible that luck will indeed affect labor supply because it affects the chances of winning the prize, and therefore, the returns to effort. Unlucky workers who have very low chances of winning may decide to leave earlier than luckier workers. But very lucky workers might expect to win with certainty, which might induce them to stop working earlier. We therefore expect a concave relationship between workers' luck (i.e., the fraction of screens displayed that included a gold square) and labor supply.[2] The estimates in columns 1 and 2 of Appendix Table C1 confirm this prediction using an expanded version of the regressions of Table 1 (columns 1 and 2) with controls for the average of the fraction of gold squares (and square of that fraction) offered to workers in the room. We also confirm this at the worker level in the last two columns of Appendix Table C1, clustering standard errors at the room level. Across all of these models, controlling for luck has the dual effect of increasing the point estimate for TP15 and decreasing its variance, rendering the estimated lower labor supply in TP15 (compared to TP30) statistically significant.[3]

These results, together with the imprecise estimates from the main model support the interpretation that effort is increased by higher tournament prizes. Theory is less clear about the effects of prize amount on labor costs. Even if TP15 induced slightly lower effort than TP30, it was still possible for labor costs to be lower because of the smaller prize amount. That was not the case. All the coefficients on TP15 in Table 2 are positive and insignificant, indicating insignificantly higher costs relative to TP30.

C.2 Comparison between TP30 and PR

We set the PR amount to match average employer costs per point in TP30, but this is not the same as matching the marginal benefit of effort for workers in the two treatments. In particular, workers' payoffs in TP30, but not in PR, depend on the behavior of their competitors. For workers in TP30 who expect to face 3 other competitors for the full 40 minutes, the expected bonus payment for working the full time is close to the expected value of the incremental bonus

[2] If workers decided to leave after reaching a certain number of points (10.7% of workers in PR and 20% of workers in FP reported setting a point goal, in the tournament rooms, one worker, in TP30, set such a goal) then we would expect to find negative (or concave) relationships between work time and luck. In PR, for lucky workers the income effect may become larger than the substitution effect. In that case, they will decide to leave, and we expect concave relationships.

[3] The results remain intact if we restrict attention to the tournament treatments, run OLS regressions for worktime, or add demographic background characteristics to the regressions.

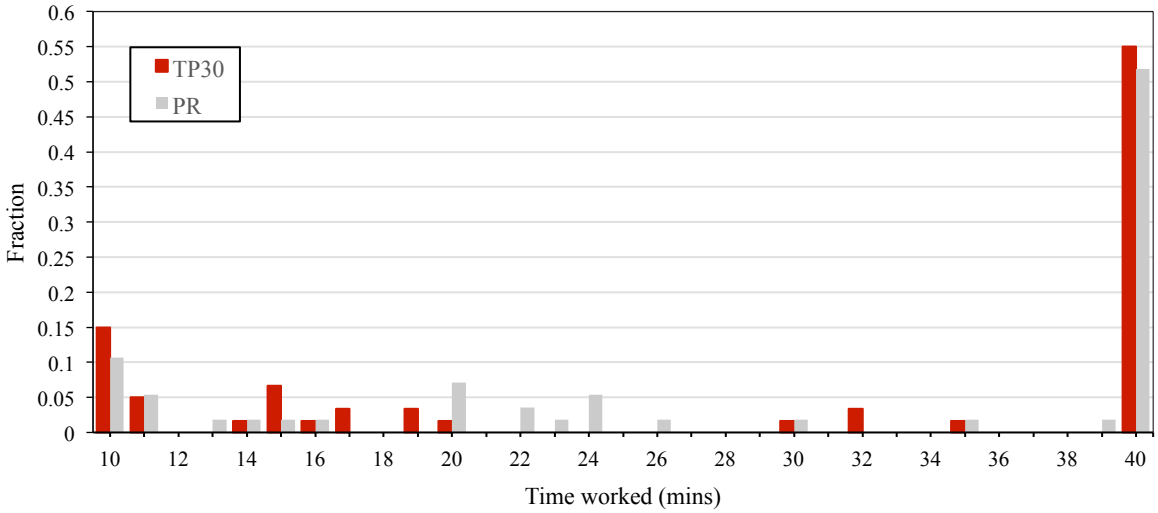
earned in in PR for staying an extra 30 minutes. The TP30 worker who works 40 minutes along with 3 others in the room should anticipate a 25% probability of winning the tournament, which implies an expected bonus of $7.50. Assuming the average cycle of active and rest screens lasts 11.5 seconds, workers should expect to see 156 red squares and 15 gold squares over the course of 30 minutes of (optional) work time. At the PR price, this yields an expected (incremental) bonus of $7.80. This similarity suggests that comparable shares of (risk neutral) workers might be willing to work the full time in PR and TP30. Indeed, as shown in Table 1, we received similar average effort from workers in the two treatments.

We also examined the prediction from the war of attrition setup that workers in TP30 should be more likely to leave the room after the other 3 competitors have left. In a pure war of attrition, the last TP30 worker in a room has no incentive to work more. Our setting has luck, and coworker output is unobservable, so it could still be worthwhile to supply some additional effort, but that is unlikely to mean working for much longer. Without a fellow competitor at work, we do not expect TP30 workers to stay the full 40 minutes, so rooms with only one worker staying 40 minutes should be relatively rare in TP30. There is no such expectation in PR, where compensation is set by individual performance. Consistent with this prediction, we find at least two workers at the end of the session in all 13 of the TP30 rooms with at least one worker persisting for the full 40 minutes. In contrast, 4 out of the 12 PR rooms in which at least one worker worked the full 40 minutes had only one person working the full time. This difference across treatments is statistically significant (Fisher exact test yields $p = 0.039$). We also formalized this with a regression model in which the outcome is the difference in quit times between the last 2 workers in the room. We find that this value is significantly smaller in TP30 as compared to PR.

Finally, we considered the stronger prediction from the finite horizon war of attrition setting that individuals will decide from the outset to either compete for the prize and work to the end or to leave immediate and incur no additional effort costs. Although this is what we find for the large majority of our sample, as shown in Figure 1 in the main text and in Appendix Figure C1, it is not universal: about 30% worked between 11 and 39 minutes. These individuals might have been taking additional time to see what others in the room were doing; or they might have decided not to compete, but stayed longer than the minimum because of the behavioral motivations described in Section 2.1.2. Even without behavioral or peer considerations, the PR

setting lacks a stark prediction about workers being driven to the endpoints of the available time, though that would be implied if both the costs of effort and the utility of income were linear over the work period. In that case, low effort cost workers would stay until the end and high cost workers would leave at 10 minutes. In typical settings with continued employment, it is more natural to expect effort costs to be convex and increasing in work hours and for the marginal utility of income (or consumption goods) to be declining. These forces would lead to more dispersion in work time in PR as compared to tournament settings, because only PR workers would find it worthwhile to invest intermediate or moderate numbers of work hours. We see a hint of this in our sample, as depicted in Appendix Figure C1: 37.5% of workers in PR left between 11 and 39 minutes. This corresponds to a 25% increase over the rate in TP30, consistent with tournament structure polarizing work times, though the difference between the two treatments is not statistically significant.

Figure C1: Histogram of Time Worked in TP30 and PR



Appendix Table C1: Effects of Treatments and Luck on Labor Supply

	(1)	(2)	(3)	(4)
	Room Level		Worker Level	
	Time Worked	Total Effort in the Session	Time Worked	Total Effort in the Session
		Main Treatments		
FP	-12.031***	-63.227***	-17.681***	-64.268***
	[2.437]	[12.567]	[3.343]	[11.857]
		Auxiliary Treatments		
TP15	-5.243*	-25.920*	-7.486*	-23.834*
	[2.711]	[13.907]	[4.162]	[13.675]
PR	-0.294	-5.700	-1.217	-3.851
	[2.532]	[11.780]	[4.009]	[11.697]
		Luck		
Fraction Gold	653.049***	3,337.135***	701.906***	2,623.025***
	[195.432]	[1,067.813]	[144.688]	[501.627]
Fraction Gold Squared	-2,993.232**	-15,144.249**	-3,471.844***	-12,916.004***
Constant	-2.071	-13.233	4.450	31.676
	[7.789]	[43.945]	[6.032]	[21.841]
Cluster Level	None	None	Room	Room
Prob>F: TP15 = FP	0.018	0.019	0.007	0.006
Prob>F: PR = FP	< 0.001	< 0.001	0.201	0.405
Prob>F: TP15 = PR	0.086	0.163	0.143	0.149
Observations	59	59	236	236
R^2		0.483		0.269

Notes: Observations at the room level in Columns 1 and 2 and worker level in Columns 3 and 4. Tobit model is estimated for time worked and linear regressions for effort measures. Fraction gold is the fraction of active screens displayed that included a gold square. Robust standard errors in brackets. *** p < 0.01, ** p < 0.05, * p < 0.1